软件学报 ISSN 1000-9825, CODEN RUXUEW
2025,36(7):3405−3428 [doi: 10.13328/j.cnki.jos.007287] [CSTR: 32375.14.jos.007287]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

基于自适应剪枝的满足本地差分隐私的真值发现算法*

张朋飞[1], 朱伊波[1], 程 祥[2,3], 张治坤[4], 刘西蒙[5], 孙 笠[6], 方贤进[1], 张 吉[7]

[1](安徽理工大学 计算机科学与工程学院, 安徽 淮南 232001)

[2](北京邮电大学 计算机学院 (国家示范性软件学院), 北京 100876)

[3](网络与交换技术国家重点实验室 (北京邮电大学), 北京 100876)

[4](浙江大学 计算机科学与技术学院, 浙江 杭州 310058)

[5](福州大学 计算机与大数据学院, 福建 福州 350108)

[6](华北电力大学 控制与计算机工程学院, 北京 102206)

[7](School of Mathematics, Physics and Computing, University of Southern Queensland, Toowoomba 4350, Australia)

通信作者: 方贤进, E-mail: xjfang@aust.edu.cn

摘 要: 为了对移动群智感知中工人上传的不同质量的感知数据做必要的聚合处理, 真值发现技术应运而生, 其是为后续应用提供精确数据支持的基础. 为了应对可能的隐私泄露问题, 现有研究往往结合本地差分隐私技术来进行保护, 然而这些研究往往忽略了感知数据中的异常值对本地差分隐私下真值发现精度的影响. 这些异常值往往具有极大的取值范围, 导致注入数据中的噪音量较大. 而且在现实世界中, 工人出于对隐私泄露的担心, 移动群智感知服务器无法在无隐私保护的情况下预先处理数据. 为解决以上问题, 提出基于自适应剪枝的满足本地差分隐私的真值发现算法 NATURE. 该算法的核心思想是考虑数据中蕴含的噪音类型来自适应剪枝掉不需要的工人的所有值或者某些任务值. 在 NATURE 中, 为便于剪枝, 在形式化约束优化问题的基础上, 设计基于优化问题的噪音感知的权重和重要性估计方法; 为进行剪枝, 在证明最优剪枝问题是 NP-hard 的基础上, 设计具有多项式时间复杂度的效用感知的自适应剪枝方法. 进一步从理论上分析 NATURE 的隐私、效用和复杂度. 在两个真实数据集和一个合成数据集上的实验结果表明, 相较于对比算法, NATURE 在求得噪音"真值"的精度上至少提高 20%.

关键词: 移动群智感知; 真值发现; 隐私保护; 本地差分隐私; 自适应剪枝

中图法分类号: TP309

中文引用格式: 张朋飞, 朱伊波, 程祥, 张治坤, 刘西蒙, 孙笠, 方贤进, 张吉. 基于自适应剪枝的满足本地差分隐私的真值发现算法. 软件学报, 2025, 36(7): 3405–3428. http://www.jos.org.cn/1000-9825/7287.htm

英文引用格式: Zhang PF, Zhu YB, Cheng X, Zhang ZK, Liu XM, Sun L, FAng XJ, Zhang J. Locally Differentially Private Truth Discovery Algorithm via Adaptive Pruning. Ruan Jian Xue Bao/Journal of Software, 2025, 36(7): 3405–3428 (in Chinese). http://www.jos.org.cn/1000-9825/7287.htm

Locally Differentially Private Truth Discovery Algorithm via Adaptive Pruning

ZHANG Peng-Fei[1], ZHU Yi-Bo[1], CHENG Xiang[2,3], ZHANG Zhi-Kun[4], LIU Xi-Meng[5], SUN Li[6], FANG Xian-Jin[1], ZHANG Ji[7]

[1](School of Computer Science and Engineering, Anhui University of Science and Technology, Huainan 232001, China)

[2](School of Computer Science (National Pilot Software Engineering School), Beijing University of Posts and Telecommunications, Beijing 100876, China)

* 基金项目: 安徽理工大学高层次引进人才科研启动基金 (2023yjrc92); 国家自然科学基金面上项目 (62372051); 国家自然科学基金青年项目 (62202164); 安徽省科技重大专项 (18030901025); 高校学科 (专业) 拔尖人才学术资助项目 (gxbjZD2021050)
收稿时间: 2023-11-29; 修改时间: 2024-04-11, 2024-08-07; 采用时间: 2024-08-30; jos 在线出版时间: 2025-01-16
CNKI 网络首发时间: 2025-01-17

[3](State Key Laboratory of Networking and Switching Technology (Beijing University of Posts and Telecommunications), Beijing 100876, China)

[4](College of Computer Science and Technology, Zhejiang University, Hangzhou 310058, China)

[5](College of Computer and Data Science, Fuzhou University, Fuzhou 350108, China)

[6](School of Control and Computer Engineering, North China Electric Power University, Beijing 102206, China)

[7](School of Mathematics, Physics and Computing, University of Southern Queensland, Toowoomba 4350, Australia)

Abstract: To conduct necessary aggregation on varying-quality sensed data uploaded by workers in mobile crowdsensing, truth discovery technology has emerged as the cornerstone for providing precise data support for subsequent applications. Existing studies tend to adopt local differential privacy for protection against potential privacy breaches, but often ignore the influence of outliers in the sensed data on the truth discovery accuracy under local differential privacy. These outliers often have a large range of values, resulting in a large amount of noise in the injected data. Additionally, due to workers' concerns about privacy breaches, mobile crowdsensing servers cannot preprocess data without privacy protection. To this end, this study proposes NATURE, which meets local differential privacy based on adaptive pruning. The core idea of the algorithm is to consider the noise types in the data to adaptively prune all unnecessary workers' values or certain task values. In NATURE, the noise-aware weight and importance estimation (NWIE) method based on a formalized constraint optimization problem is designed to facilitate data pruning. Based on proving the optimal pruning problem is NP-hard, this study designs the utility-aware adaptive pruning (UAP) method with polynomial time complexity to conduct pruning. Furthermore, a theoretical analysis of NATURE's privacy, utility, and complexity is carried out. Experimental results on two real-world datasets and one synthetic dataset demonstrate that NATURE achieves an accuracy improvement of at least 20% in obtaining "truth" compared to its comparative algorithms.

Key words: mobile crowdsensing (MCS); truth discovery; privacy protection; local differential privacy; adaptive pruning

随着智慧城市建设步伐的加快, 环境感知需求日益增长并呈现复杂化趋势, 移动群智感知 (mobile crowdsensing, MCS) 技术的出现给复杂环境下的城市感知带来了更精确有效的解决方案[1]. 通过工人间的协作, 可为诸如环境监测、罪犯追踪以及智慧交通等提供决策支持[2]. 相关统计显示, MCS 的部署可以降低 30%–70% 的交通事故, 减少不必要的生命财产损失[3]. 在 MCS 中, 感知数据的收集被外包给了一大群携带感知设备的工人, 这些工人享受 MCS 提供的服务并向 MCS 服务器提供数据以改善 MCS 的后续服务[4].

在 MCS 中, 不同工人设备的感知精度不同, 同时由于网络拓扑结构的影响, 数据传输环境也可能不可靠, 因此, MCS 服务器需要将收集到的数据进行必要的聚合处理[5]. 传统的解决方案是对感知数据求平均值. 然而, 因为该方法平等地对待所有工人的数据, 可能导致无法得出准确的结果. 因为不同工人所感知到的数据质量差异较大, 而且每个工人的权重通常是未知的[6], 理想的方法应该具有捕捉差异的能力. 为应对这一挑战, 旨在从众多质量不同的数据中发现"真值"的真值发现方法应运而生. 该方法的核心思想是如果一个工人的数据更接近"真值", 则给该工人分配更高的权重[7].

然而, 工人上传的数据中可能隐含个人的隐私信息, 比如家庭住址、身份信息等可能从上传的数据中推导出来[8,9]. 因此, 如果不加处理地将这些数据上传到服务器可能导致工人因为担心隐私泄露而不愿意参与到 MCS 中来, 进而影响 MCS 系统的运行和用户对 MCS 的粘性. 为避免隐私泄露, 研究人员提出使用本地差分隐私 (local differential privacy, LDP) 技术来保护工人的隐私[10], 其核心思想是数据在上传到服务器之前在本地进行扰动, 能从根本上保护用户的隐私. 目前, LDP 技术已被应用于众多互联网巨头的产品中, 比如谷歌的 Chrome 浏览器[11]、苹果的 iOS 系统[12]以及微软的 Windows Insiders[13]. 一种快捷、高效的实现 LDP 的典型方法是向数据中注入 Laplace 噪音[14]. 本文也将采用 Laplace 来保护工人的数据.

在真值发现场景中, 由于主客观因素的影响, 每个工人上传的感知数据质量可能显著不同, 有些数据甚至不可用. 从主观上来说, 实际的 MCS 系统中可能存在恶意的工人, 提交错误的值. 也可能存在女巫攻击者 (sybil attacker) 故意提交精心构造的其他值来恶意提升自己的感知质量或者是获得 MCS 服务器的奖励[15]; 从客观上来说, 有些工人可能并不知道自身传感器已失效, 也可能受周围环境以及传输网络的影响[16]; 非隐私下真值发现算法也需要对数据做预处理[6,7]. 真值发现算法对不同数据加权聚合, "真值估计"步骤和"权重更新"步骤紧密耦合, 在没有任何先验知识的情况下, 由于错误甚至是恶意数据的干扰, 可能导致得到的"真值"精度低或者需要

较高的迭代次数. 现有研究表明, 即便对这些质量很低的数据赋予消极的权重, 也会对最终结果产生较大的负面影响[14,17−21].

现有满足 LDP 的真值发现方法[22−27]忽略了数据中的异常值对本地差分隐私下真值发现精度的影响, 这些异常值往往具有极大的取值范围, 导致注入数据中的噪音量较大. 而且实际中工人出于对隐私泄露的担心, 移动群智感知服务器无法在非隐私下预先处理. 一种直观方案是在满足 LDP 下采样一些需要的值或者剪枝掉不需要的值, 并根据加噪上传后的值进行真值发现. 然而, 采样比例越大注入的噪音就越大, 采样比例越小剩下数据中的信息量就越小, 难以确定一个恰当的采样比例. 而且, 该采样比例的确定很容易受到异常值的干扰. 此外, 从剪枝掉的数据量来说, 如果剪枝掉的数据量太少, 则无法起到剪枝的效果, 剪枝太多可能导致剩下数据的信息量不够; 从剪枝掉的工人来说, 这些工人可能是高质量的工人, 这会降低留下数据的质量; 从剪枝掉的任务值来说, 随机剪枝可能导致某些任务没有值. 这是因为某些任务可能被少量工人做过, 如果这些值都被剪枝掉, 那么就无法得到这些任务对应的“真值”, 进而导致真值发现的失败.

为此, 本文从剪枝的角度, 提出一种基于自适应剪枝的满足 LDP 的真值发现算法 (locally differentially private truth discovery via adaptive pruning, NATURE). 该算法的核心思想是考虑数据中蕴含的噪音类型效用感知地自适应剪枝掉不需要的工人的所有值或者某些任务值, 然后利用剩下的噪音值进行真值发现. 具体地讲, 首先每个工人采用 Laplace 机制对自身数据加噪并将加噪后数据上传到群智感知服务器, 其次为便于剪枝, 本文在形式化约束优化问题的基础上求得工人的质量和任务的重要性以及每个任务的初步“真值”; 接着, 为进行剪枝, 本文将剪枝问题形式化为另一个约束优化问题并证明该问题是 NP-hard 的, 且进一步根据工人的质量、任务的重要性以及算法的总误差设计了具有多项式时间复杂度的、效用感知的自适应剪枝方法; 最后, 根据剪枝后的噪音数据和任务的初步“真值”进行非隐私的真值发现以得到每个任务的最终噪音“真值”.

综上, 本文的主要贡献如下.

(1) 基于剪枝思想, 本文提出一种针对异常值的满足 LDP 的真值发现算法 NATURE, 并从理论上分析了算法的隐私、效用和复杂度. 理论分析表明 NATURE 不仅严格满足 LDP, 而且还能得到高精度的噪音“真值”, 同时相对于非隐私的真值发现算法来说, 没有增加多余的通信开销, 计算开销也增加较少.

(2) 为便于剪枝, 本文考虑上传数据中蕴含的噪音类型, 形式化了约束优化问题, 并基于拉格朗日乘子法来求解得到工人的质量和任务的重要性以及任务的初始“真值”.

(3) 为进行剪枝, 本文将剪枝问题形式化为另一个约束优化问题并证明该问题是 NP-hard 的, 进而基于工人的质量和任务的重要性设计了具有多项式时间复杂度的、效用感知的自适应剪枝算法.

(4) 本文在两个真实的数据集和一个合成数据集上验证了 NATURE 的效用和效率. 相较于对比算法, NATURE 在求得噪音“真值”的精度上至少提高 20%.

1 相关工作

本文相关工作主要落在本地差分隐私 (LDP) 中的数据投毒与防御, 基于密码学的真值发现和满足 LDP 的真值发现.

1.1 LDP 中的数据投毒与防御

Cao 等人[28]和 Cheu 等人[29]指出 LDP 协议对数据操纵较为敏感, 攻击者通过注入精心制作的虚假数据以使得 LDP 协议错误估计项目的频率或者均值, 此种攻击被称为“数据投毒攻击 (data poisoning attack, DPA)”. Li 等人[30]为了平衡群智感知系统中的安全风险与隐私威胁, 提出一种新的基于伪装的 DPA 方法, 来欺骗满足差分隐私的真值发现协议.

为防御 DPA, Zhang 等人[31]提出了一种多轮次的针对真值发现的攻击和防御方法. Song 等人[32]提出了两种可验证的 LDP 协议, 优化了现有使用零知识证明协议存在计算和通信成本高的问题. Huang 等人[33]针对频率估计等数据分析任务提出了通用的 LDP 投毒防御机制. 在此基础上, Zheng 等人[34]针对较低隐私预算下基于 LDP 的

群智感知更容易受到 DPA 的问题, 设计了基于差分进化算法的梯度代替方法.

与此同时, 研究者们也往往面向联邦学习场景设计投毒攻击和防御方法. 为了防御模型投毒攻击 (model poisoning attack, MPA), Hossain 等人[35]基于强化学习为本地客户模型选择不同的隐私级别, 从而平衡模型泄露和集成模型的效用. Zhong 等人[36]基于拜占庭鲁棒设计了一种基于统计的防御方法, 以消除特征层面的异常干扰. 进一步地, Huang 等人[37]通过服务器验证被差分隐私噪音扰动的模型参数, 检测发现集中模型中的有毒参数和恶意用户.

尽管 LDP 协议中的投毒攻击与防御问题和本文研究均是为了获取正确的统计信息, 然而二者的问题场景和采用技术却有较大区别. 就问题场景而言, 投毒攻击和防御是针对“恶意值”的场景, 其中部分工人是恶意的, 其采用 LDP 技术是为了隐藏自身的恶意提交值, 可能并不会严格遵守 LDP 协议. 同时, 这些提交这些“恶意值”是为了破坏服务器收集的统计信息, 这些值可能并不具备较大取值范围, 也就是说投毒防御的目的是识别精心构造的虚假值. 与之不同的是, 本文针对的是“异常值”场景, 其中工人会严格遵守 LDP 协议, 提交值具有大的取值范围, 本研究是为了在满足 LDP 保护的注入噪音就较大情况下得到精确的任务“真值”. 就采用技术而言, 投毒攻击与防御往往需要攻击者与防御者具有一定背景知识, 比如投毒者的数量, 或者操纵的一系列的其他工人的提交值. 与之不同的是, 本文同时考虑为满足 LDP 保护注入的拉普拉斯噪音和表示工人质量的高斯噪音, 在不需要背景知识的情况下自适应剪枝掉“异常值”并获得最终任务的“真值”.

1.2 基于密码学的真值发现

Zhang 等人[38]针对不同场景提出了两种可靠的真值发现方法: 第 1 种是针对工人位置相对稳定的场景, 他们应用 Paillier 同态加密、单向哈希链和超增序列等技术来实现隐私保护, 第 2 种是针对工人经常移动的场景主要采用 Paillier 同态加密来实现隐私保护. Xue 等人[39]提出了一种基于激励机制的隐私保护真值发现方法, 其结合同态加密系统在两个非串谋服务器下, 引入基于权重的激励机制, 既保证了更强的安全性能, 又确保计算效率和性能. Zhao 等人[40]提出一种基于边缘计算范例的轻量级、健壮的匿名机制. 该方法通过异常值检测来消除异常值的影响, 以实现真值发现结果的鲁棒性. Wu 等人[41]提出了两种真值发现方法: 一种是基础隐私感知真值发现方法, 另一种是具有两台服务器的隐私增强真值发现方法, 同时兼顾隐私性和实用性. Tang 等人[42]提出了匿名化真值发现方法, 将工人与他们的数据断开链接, 其中工人的数据无需复杂的加密即可计算和传输, 还借助扰动技术提出了一种更轻量级的真值发现方法, 以在工人愿意公开权重的场景中进一步减少每个工人的开销. Zhang 等人[43]基于矩阵计算特性, 设计了基于密钥推导和 (重新) 加密机制的新方法, 以在真值发现过程中同时保护用户的任务隐私和数据隐私. Liu 等人[44]提出一种协作式云加密的隐私保护真值发现系统架构, 该架构在利用乱码电路的同时适应双云对等模型. Xiong 等人[45]基于秘密共享和加性同态特性提出一种去中心化的隐私保护框架. Liu 等人[46]进一步结合数据掩码来解决时空关联下的真值发现问题.

由于本文采用 LDP 来保护工人数据, 并不涉及任何密码学相关内容, 因此这些优秀成果和本文的研究内容正交.

1.3 满足 LDP 的真值发现

国内外针对满足 LDP 的真值发现研究尚处于起步阶段. 现有研究主要针对不同的数据类型展开满足 LDP 的真值发现工作.

针对离散值的情况, Sun 等人[24]基于随机响应机制提出一种面向个性化激励的真值发现算法, 然而该算法只能适用于二值属性. Li 等人[47]提出一种分层的 LDP 的真值发现算法. 其中首先从超分布中采样一个概率, 然后基于此概率调用随机响应机制来扰动工人数据. Wang 等人[48]提出一种仅保护部分属性的新方法.

针对连续值的情况, Sun 等人[22]针对稀疏众包场景, 首先定义了非严格的 LDP 范式 Cell-LDP, 在此基础上提出了基于矩阵分解的新方法. 与文献 [47] 类似, Li 等人[23]结合高斯机制提出了分层的真值发现方法. 由于采用高斯机制, 该方法仅满足 (ε,δ)-LDP. 为了处理个性化激励的场景, Sun 等人[25]首先定义了非严格的 LDP 范式 KL-

LDP, 在此基础上基于拍卖方法设计了满足个体理性等激励要求的真值发现新方法. 为了严格满足 LDP, Zhang 等人[26]结合采样与推断提出了一种针对高维场景的严格满足 LDP 的新方法. 进一步地, 他们[27]结合概率化估计与聚合来提升现有方案的精度.

综上, 现有研究要么针对离散值所设计, 要么对于连续值不严格满足 LDP. 同时现有研究均潜在假设数据是纯净的或者是能在非隐私下被预先处理的. 然而实际中工人出于隐私泄露的担心, 群智感知服务器无法在非隐私下预先处理数据. 为此, 本文针对连续值提出基于自适应剪枝的新方法 NATURE.

2 预备知识

2.1 真值发现

真值发现的典型算法是 CRH (conflict resolution on heterogeneous data)[7]. 其遵循真值发现的一般原则, 即给予数据更接近“真值”的工人更高的权重, 通过对来自不同质量工人提供的数据进行加权聚合来得到最终的“真值”. 通过交替迭代“真值估计” (公式 (1)) 和“权重更新” (公式 (2)) 直至收敛.

$$x_n^* = \frac{\sum_{s=1}^{M}\sum_{n=1}^{N} w_s \cdot x_n^s}{\sum_{s=1}^{M}\sum_{n=1}^{N} w_s} \tag{1}$$

$$w_s = -\log\left(\frac{\sum_{n=1}^{N} d(x_n^s, x_n^*)}{\sum_{s=1}^{M}\sum_{n=1}^{N} d(x_n^s, x_n^*)}\right) \tag{2}$$

其中, M、N 分别是工人数和任务数. w_s 是第 s 个工人的权重, x_t^s 是第 s 个工人提交给第 t 个任务的值.x_t^* 是第 t 个任务计算得到的真值, $d(\cdot)$ 往往采用欧氏距离来估算工人提交值和估算真值的差距.

2.2 本地差分隐私

本地差分隐私 (LDP) 的形式化定义如下.

定义 1. LDP. 假设有一个随机算法 $\mathcal{M}$, 其定义域为 $Dom(\mathcal{M})$, 值域为 $Ran(\mathcal{M})$, 如果对于任意两条数据记录 t、$t' \in Dom(\mathcal{M})$, 以及任意一个输出 $t^* \in Ran(\mathcal{M})$, 有下列不等式成立:

$$\Pr[\mathcal{M}(t) = t^*] \leqslant \mathrm{e}^{\varepsilon} \cdot \Pr[\mathcal{M}(t') = t^*],$$

则算法 $\mathcal{M}$ 满足 ε-LDP, 其中 ε 为隐私保护参数.

对于有些复杂的隐私保护问题, 通常需要多次应用 LDP 算法, 此时需要将隐私预算进行合理分配.

定理 1. 串行组合定理. 假设有满足 LDP 的随机算法 $\mathcal{M}_1, \mathcal{M}_2, \ldots, \mathcal{M}_n$, 其对应的隐私参数分别为 $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$, 当在同一数据集上使用这些随机算法时, 它们构成的算法提供 $\left(\sum_{i=1}^{n} \varepsilon_i\right)$-LDP.

定理 2. 并行组合定理. 假设有满足 LDP 的随机算法 $\mathcal{M}_1, \mathcal{M}_2, \ldots, \mathcal{M}_n$, 其对应的隐私参数分别为 $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$, 当这些随机算法作用于不相交的数据集 $D_1, D_2, \ldots, D_n$ 时, 它们构成的算法对这些数据集提供 $\max(\varepsilon_i)$-LDP.

定理 3. 后处理定理. 假设有满足 LDP 的随机算法 $\mathcal{M}_A$ 和工作在 $\mathcal{M}_A$ 输出上的任意算法 $\mathcal{M}_B$, 那么算法 $\mathcal{M}_B$ 也满足同级别隐私保护强度的 LDP.

一种典型的方法是采用 Laplace 机制, 在该机制中向数据中的每一维注入服从均值为 0 的 Laplace 分布的噪音 $f(x) = \frac{1}{2\lambda}\mathrm{e}^{-\frac{|x|}{\lambda}}$. 其中 $\lambda = \frac{\Delta f}{\varepsilon}$, Δf 是查询函数的敏感度. 比如如果查询某个值, 该值的范围为 $[d_1, d_2]$, 那么 $\Delta f = d_2 - d_1$.

2.3 问题陈述

如图 1 所示, 假设有 N 个任务 $T = \{t_1, t_2, \ldots, t_N\}$ 需要进行真值发现, 共有 M 个工人 $U = \{u_1, u_2, \ldots, u_M\}$ 参与到 LDP 的真值发现中来. 工人 s 做了 $|M_s|$ 个任务. 每个工人的隐私预算为 ε. 本文的目标是在严格满足 LDP 的条件下保护各个工人的值 x_t^s, 求得各个任务的真值 $\hat{x}_n^*$.

如表 1 所示是本文可能用到的主要符号说明.

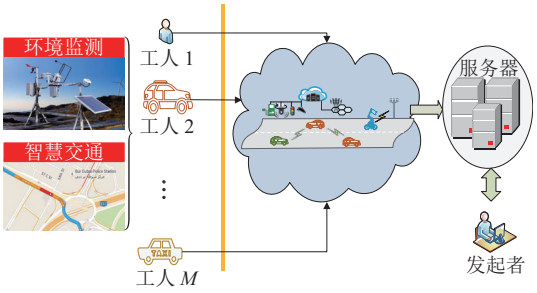


图 1 本文问题场景

表 1 主要符号

变量	变量描述
M	工人数
N	任务数
w_s	第s个工人的权重
x_n^s	第s个工人关于第n个任务的感知数据
x_n^*	第n个任务的真值
U	工人集
T	任务集
$\hat{x}_n^*$	噪音真值
$x_n^{s\prime}$	x_n^s对应的带噪值
U_n	做了第n个任务的用户集合
d_2, d_1	某任务取值的上下界
y_n	第n个任务的重要性

3 NATURE 算法

3.1 算法概述

一种直观方法是调用文献 [26] 的方法采样一部分数据来加噪并进行真值发现, 然而, (1) 该方法并不鲁棒, 很容易受到异常值的影响. (2) 未被采样到的工人可能是高质量工人. (3) 某些任务可能只被少量工人做过, 如果这些值未被采样到, 那么就无法得到这些任务对应的“真值”, 进而导致真值发现的失败.

为了提升得到的噪音“真值”效用, 本文需要去掉异常值, 从而降低这些值对“真值”精度的负面影响. 为此, 本文提出 NATURE 算法, 该算法的主要思想是考虑数据中蕴含的噪音模式来对 Laplace 机制保护后的数据自适应地剪枝之后进行非隐私的真值发现. 特别地, 在 NATURE 中, 根据工人上传的加噪数据, 本文提出一种噪音感知的方法来求得工人的质量和任务的重要性 (noise-aware weight and importance estimation, NWIE) 从而为剪枝提供支持. 由于所提方法是噪音敏感的, 因此使得 NATURE 不容易受到异常值的影响. 再者, 本文提出一种根据工人的质量、任务的重要性和 LDP 下真值发现的总体误差的效用感知的自适应剪枝方法 (utility-aware adaptive pruning, UAP) 来进行剪枝, 使得 NATURE 能保留高质量的工人和任务数据, 从而得到每个任务的“真知”且求得的噪音“真值”精度较高. 如图 2 所示, 该算法主要包含如下阶段.

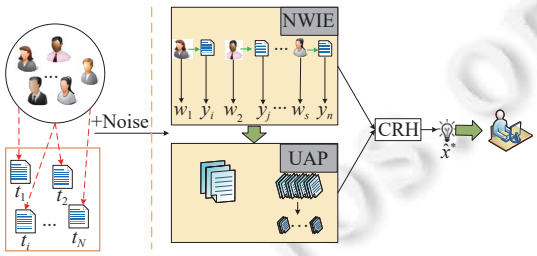


图 2 NATURE 整体流程

阶段 1. 工人调用 Laplace 机制上传加噪数据给服务器;

阶段 2. 服务器根据加噪数据调用 NWIE 估算工人的质量和任务的重要性以及任务的初步真值;

阶段 3. 服务器根据工人的质量和任务的重要性调用 UAP 进行剪枝, 从而留下高质量和对真值发现重要的数据;

阶段 4. 服务器根据剪枝后的数据和任务的初步“真值”调用 CRH 得到每个任务的噪音“真值”, 并提交给移动群智感知的发起者.

3.2 噪音感知的权重和重要性估计

在得到工人上传的噪音值 $x_n^{s\prime}$ 之后, 为了对工人和任务剪枝提供必要的依据, 除了工人的权重之外, 本文引入任务的重要性, 用于衡量工人提交的关于该任务的值对最终精度的影响. 考虑到数据中蕴含的表示工人质量的高斯噪音[6,7]和为隐私保护注入的拉普拉斯噪音, 根据噪音数据求得"真值"的建模方法如下.

在本文的模型中, 每个观测值 x' 都是一个以真实值 x^* 为中心、协方差为 $\sigma_i^2 I_N$ 的混合分布, 其中 α 是混合分布的比例. 实际中受到主客观因素的影响, 每个工人不可能是一个完美的工人, 也就是说每个工人的可靠性存在上界, 因此本文限制 $\sigma_i \geqslant \sigma_0$. 最大化以下似然函数:

$$\begin{aligned}&\prod_{i=1}^{M}\left[\alpha\mathcal{N}(x^*\mid x',\sigma_i^2 I_N)+(1-\alpha)\mathcal{L}(x^*\mid x',2\sigma_i^2 I_N)\right]\\&=\prod_{i=1}^{M}\left(\alpha\left(\frac{1}{\sqrt{2\pi}\sigma_i}\right)^N\exp\left[-\frac{\|x_i'-x^*\|^2}{2\sigma_i^2}\right]+(1-\alpha)\left(\frac{1}{2\sigma_i}\right)^N\exp\left[-\frac{\|x_i'-x^*\|}{\sigma_i}\right]\right)\end{aligned}\tag{3}$$

取负对数后展开得到:

$$\frac{MN}{2}\ln(2\pi)-M\ln\alpha+\sum_{i=1}^{M}\left(N\ln\sigma_i+\frac{\|x_i'-x^*\|^2}{2\sigma_i^2}-\ln\left(1+\frac{1-\alpha}{\alpha}\left(\frac{\sqrt{2\pi}}{2}\right)^N\exp\left[\frac{\|x_i'-x^*\|^2}{2\sigma_i^2}-\frac{\|x_i'-x^*\|}{\sigma_i}\right]\right)\right)\tag{4}$$

由于 $\ln(1+x)\leqslant x$ 且 $\mathrm{e}^x\approx 1+x$, 得到以下优化问题:

$$\min_{x^*,\sigma}\left\{\frac{MN}{2}\ln(2\pi)-M\ln\alpha+\sum_{i=1}^{M}\left\{\begin{array}{l}N\ln\sigma_i+\dfrac{\|x_i'-x^*\|^2}{2\sigma_i^2}-\dfrac{1-\alpha}{\alpha}\left(\dfrac{\sqrt{2\pi}}{2}\right)^N\\\left(\dfrac{\|x_i'-x^*\|^2}{2\sigma_i^2}-\dfrac{\|x_i'-x^*\|}{\sigma_i}+1\right)\end{array}\right\}\right\},\ \text{s.t. }\sigma_i\geqslant\sigma_0\tag{5}$$

由于 $\sigma_i=\max\left\{\sigma_0,\|x_i'-x^*\|/\sqrt{N}\right\}$, 同时根据方差公式并便于计算, 定义 $\sigma_0=1/\sqrt{N}$, 得到:

$$\min_{x^*}\left\{\sum_{\|x_i'-x^*\|<1}\left(\frac{d\|x_i'-x^*\|^2}{2}-\frac{1-\alpha}{\alpha}\left(\frac{\sqrt{2\pi}}{2}\right)^N\left(\frac{N}{2}\|x_i'-x^*\|^2-\sqrt{N}\|x_i'-x^*\|\right)\right)+\sum_{\|x_i'-x^*\|\geqslant 1}(N\ln\|x_i'-x^*\|)\right\}\tag{6}$$

假设 $\ell=\|x_i'-x^*\|$, 最终的优化问题为:

$$\left\{\begin{array}{ll}\left[1-\dfrac{(1-\alpha)}{\alpha}\left(\dfrac{\sqrt{2\pi}}{2}\right)^N\right]\ell^2-\dfrac{2\sqrt{d}(1-\alpha)}{d\alpha}\left(\dfrac{\sqrt{2\pi}}{2}\right)^N\ell, & 0\leqslant\ell<1\\2\ln\ell, & \ell\geqslant 1\end{array}\right.\tag{7}$$

分别用 w_s 和 y_n 表示第 s 个工人的质量和第 n 个任务的重要性. 为了同时得到工人质量和任务的重要性, 本文提出噪音感知的权重和重要性估计算法 (noise-aware weight and importance estimation, NWIE). 该算法的主要思想是同时考虑数据中蕴含的内源性的高斯噪音以及注入的外源性的 Laplace 噪音来推断工人的质量和任务的重要性.

在 NWIE 中, 根据公式 (7) 构建如公式 (8) 所示的约束优化问题. 其最小化工人上传噪音值 $x_n^{s\prime}$ 和待求解"真值" $\hat{x}_n^*$ 之间的加权距离. 同时, 考虑工人的质量和任务的重要性, 如果求得的"真值"偏离高质量的工人和重要的任务, 那么在目标函数中给予其更大的惩罚. 公式 (8) 中的两个约束条件分别用来约束权重和工人的质量分布取值在 0–1 之间.

$$\left\{\begin{array}{ll}\min\limits_{w_s,\hat{x}_n^*,y_n}\displaystyle\sum_{s=1}^{M}\sum_{n=1}^{N}w_s y_n\left[\left(1-\left(\frac{\sqrt{2\pi}}{2}\right)^N\right)(x_n^{s\prime}-\hat{x}_n^*)^2-\frac{2\sqrt{N}}{N}\left(\frac{\sqrt{2\pi}}{2}\right)^N\left|x_n^{s\prime}-\hat{x}_n^*\right|+2\ln\left|x_n^{s\prime}-\hat{x}_n^*\right|\right]\\\text{s.t.}\quad\left\{\begin{array}{l}\displaystyle\sum_{s=1}^{M}\mathrm{e}^{-w_s}=1\\\displaystyle\sum_{n=1}^{N}\mathrm{e}^{-y_n}=1\end{array}\right.\end{array}\right.\tag{8}$$

由于实数空间上的任意范数都是凸函数, 同时非负加权和以及复合仿射变换是凸保留算子, 因此上述问题是凸优化问题. 可以结合拉格朗日乘子法来设计求解方案, 具体步骤如下.

首先得到公式 (8) 对应的拉格朗日函数, 如公式 (9) 所示. 其中 λ 和 η 是拉格朗日乘子.

$$L(w_s,\hat{x}_n^*,y_n,\lambda,\eta)=\sum_{s=1}^{M}\sum_{n=1}^{N}w_s y_n\left[\left(1-\left(\frac{\sqrt{2\pi}}{2}\right)^N\right)(x_n^{s\prime}-\hat{x}_n^*)^2-\frac{2\sqrt{N}}{N}\left(\frac{\sqrt{2\pi}}{2}\right)^N\left|x_n^{s\prime}-\hat{x}_n^*\right|+2\ln\left|x_n^{s\prime}-\hat{x}_n^*\right|\right]$$
$$+\lambda\left(\sum_{s=1}^{M}\mathrm{e}^{-w_s}-1\right)+\eta\left(\sum_{n=1}^{N}\mathrm{e}^{-y_n}-1\right) \tag{9}$$

令 $t=1-\left(\frac{\sqrt{2\pi}}{2}\right)^N$, $b=\frac{2\sqrt{N}}{N}\left(\frac{\sqrt{2\pi}}{2}\right)^N$, 接着对公式 (9) 求得关于 w_s 的一阶导并令其等于 0, 可以得到公式 (10):

$$\lambda\mathrm{e}^{-w_s}=\sum_{n=1}^{N}y_n t(x_n^{s\prime}-\hat{x}_n^*)^2-y_n b\left|x_n^{s\prime}-\hat{x}_n^*\right|+2y_n\ln\left|x_n^{s\prime}-\hat{x}_n^*\right| \tag{10}$$

根据公式 (8) 中的约束条件 1, 有:

$$\lambda(\mathrm{e}^{-w_1}+\ldots+\mathrm{e}^{-w_s}+\ldots+\mathrm{e}^{-w_M})=\lambda \tag{11}$$

进而求得:

$$\lambda=\sum_{s=1}^{M}\sum_{n=1}^{N}y_n t(x_n^{s\prime}-\hat{x}_n^*)^2-y_n b\left|x_n^{s\prime}-\hat{x}_n^*\right|+2y_n\ln\left|x_n^{s\prime}-\hat{x}_n^*\right| \tag{12}$$

将求得的 λ 回代入公式 (10), 可以得到:

$$w_s=-\ln\frac{\sum_{n=1}^{N}y_n t(x_n^{s\prime}-\hat{x}_n^*)^2-y_n b\left|x_n^{s\prime}-\hat{x}_n^*\right|+2y_n\ln\left|x_n^{s\prime}-\hat{x}_n^*\right|}{\sum_{s=1}^{M}\sum_{n=1}^{N}y_n t(x_n^{s\prime}-\hat{x}_n^*)^2-y_n b\left|x_n^{s\prime}-\hat{x}_n^*\right|+2y_n\ln\left|x_n^{s\prime}-\hat{x}_n^*\right|} \tag{13}$$

同理, 将公式 (9) 对 y_n 求导可以得到:

$$y_n=-\ln\frac{\sum_{s=1}^{M}w_s t(x_n^{s\prime}-\hat{x}_n^*)^2-w_s b\left|x_n^{s\prime}-\hat{x}_n^*\right|+2w_s\ln\left|x_n^{s\prime}-\hat{x}_n^*\right|}{\sum_{s=1}^{M}\sum_{n=1}^{N}w_s t(x_n^{s\prime}-\hat{x}_n^*)^2-w_s b\left|x_n^{s\prime}-\hat{x}_n^*\right|+2w_s\ln\left|x_n^{s\prime}-\hat{x}_n^*\right|} \tag{14}$$

进一步地, 将公式 (9) 对 $\hat{x}_n^*$ 求偏导可以得到:

$$-2\sum_{s=1}^{M}w_s y_n t(x_n^{s\prime}-\hat{x}_n^*)-\sum_{s=1}^{M}w_s y_n\cdot b\cdot Sgn(x_n^{s\prime}-\hat{x}_n^*)+2\sum_{s=1}^{M}w_s y_n\cdot(-1)\cdot\frac{1}{(x_n^{s\prime}-\hat{x}_n^*)}=0 \tag{15}$$

其中, $Sgn(\cdot)$ 是符号函数, 如果内部值大于 0 则返回 1, 否则返回−1; 如果等于 0 则返回 0. 那么当 $x_n^{s\prime}>\hat{x}_n^*$, 公式 (15) 化简为:

$$-2\sum_{s=1}^{M}w_s y_n t(x_n^{s\prime}-\hat{x}_n^*)-\sum_{s=1}^{M}w_s y_n\cdot b+2\sum_{s=1}^{M}w_s y_n\cdot(-1)\cdot\frac{1}{(x_n^{s\prime}-\hat{x}_n^*)}=0 \tag{16}$$

由泰勒公式 $\frac{1}{a-x}=\frac{1}{a}-\frac{1}{a^2}x+O(x)$ 可得:

$$\frac{1}{(x_n^{s\prime}-\hat{x}_n^*)}=\frac{1}{x_n^{s\prime}}-\frac{1}{(x_n^{s\prime})^2}\hat{x}_n^*+O(\hat{x}_n^*) \tag{17}$$

所以代入公式 (16) 化简为:

$$
\begin{aligned}
&-2\sum_{s=1}^{M} w_s y_n \cdot t \cdot (x_n^{s\prime} - \hat{x}_n^*) = \sum_{s=1}^{M} w_s y_n \cdot b + 2\sum_{s=1}^{M} w_s y_n \cdot \left[\frac{1}{x_n^{s\prime}} - \frac{1}{(x_n^{s\prime})^2}\hat{x}_n^* + O(\hat{x}_n^*)\right] \\
&\Leftrightarrow -2\sum_{s=1}^{M} w_s y_n \cdot t \cdot x_n^{s\prime} + 2\sum_{s=1}^{M} w_s y_n \cdot t \cdot \hat{x}_n^* = \sum_{s=1}^{M} w_s y_n \cdot b + 2\sum_{s=1}^{M} w_s y_n \cdot \frac{1}{x_n^{s\prime}} - 2\sum_{s=1}^{M} w_s y_n \cdot \frac{1}{(x_n^{s\prime})^2}\hat{x}_n^* \\
&\Leftrightarrow 2\sum_{s=1}^{M} w_s y_n \cdot t \cdot \hat{x}_n^* + 2\sum_{s=1}^{M} w_s y_n \cdot \frac{1}{(x_n^{s\prime})^2}\hat{x}_n^* = \sum_{s=1}^{M} w_s y_n \cdot \left(b + 2t \cdot x_n^{s\prime} + 2\frac{1}{x_n^{s\prime}}\right) \\
&\Leftrightarrow \hat{x}_n^* \cdot 2\sum_{s=1}^{M} w_s y_n \cdot \left(t + \frac{1}{(x_n^{s\prime})^2}\right) = \sum_{s=1}^{M} w_s y_n \cdot \left(b + 2t \cdot x_n^{s\prime} + 2\frac{1}{x_n^{s\prime}}\right) \\
&\Leftrightarrow \hat{x}_n^* = \frac{\sum_{s=1}^{M} w_s y_n \cdot \left(b + 2t \cdot x_n^{s\prime} + 2\frac{1}{x_n^{s\prime}}\right)}{2\sum_{s=1}^{M} w_s y_n \cdot \left(t + \frac{1}{(x_n^{s\prime})^2}\right)}
\end{aligned} \tag{18}
$$

同理当 $x_n^{s\prime} < \hat{x}_n^*$ 时, 得到:

$$
\hat{x}_n^* = \frac{\sum_{s=1}^{M} w_s y_n \cdot \left(-b + 2t \cdot x_n^{s\prime} + 2\frac{1}{x_n^{s\prime}}\right)}{2\sum_{s=1}^{M} w_s y_n \cdot \left(t + \frac{1}{(x_n^{s\prime})^2}\right)} \tag{19}
$$

所以综上所述,

$$
\hat{x}_n^* = \begin{cases}
\dfrac{\sum_{s=1}^{M} w_s y_n \cdot \left(b + 2t \cdot x_n^{s\prime} + 2\frac{1}{x_n^{s\prime}}\right)}{2\sum_{s=1}^{M} w_s y_n \cdot \left(t + \frac{1}{(x_n^{s\prime})^2}\right)}, & \text{if } x_n^{s\prime} > \hat{x}_n^* \\
\dfrac{\sum_{s=1}^{M} w_s y_n \cdot \left(-b + 2t \cdot x_n^{s\prime} + 2\frac{1}{x_n^{s\prime}}\right)}{2\sum_{s=1}^{M} w_s y_n \cdot \left(t + \frac{1}{(x_n^{s\prime})^2}\right)}, & \text{if } x_n^{s\prime} < \hat{x}_n^*
\end{cases} \tag{20}
$$

NWIE 算法流程如算法 1 所示.

算法 1. NWIE 算法.

输入: 工人的值 x_n^s; 迭代阈值 τ;

输出: 工人的权重 w_s; 任务的重要性 y_n; 任务真值 $\hat{x}_n^*$.

//本地执行

1) 每个工人调用 Laplace 机制对 x_n^s 加噪;
2) 工人将加噪后的值 $x_n^{s\prime}$ 上传给服务器;

//服务器执行

3) 初始化 w_s, y_n 和 $\hat{x}_n^*$;
4) Repeat
5) 根据公式 (13) 更新 w_s;
6) 根据公式 (14) 更新 y_n;
7) 根据公式 (20) 更新 $\hat{x}_n^*$;

8) Until 达到迭代次数或者满足迭代阈值 τ;

9) Return w_s, y_n 和 $\hat{x}_n^*$;

在算法 1 中, 首先工人本地加噪并上传所做任务的值 (第 1, 2 行), 然后服务器迭代更新工人质量、任务重要性以及任务的“真值” (第 3–8 行), 最后返回相关值 (第 9 行). 其中用计算所有任务前后两次“真值”差距的平均值是否小于 τ 来判断是否满足迭代阈值 τ. 具体地讲, τ 的计算方式如公式 (21) 所示:

$$\tau = \frac{\sum_{t=1}^{N} \left| \hat{x}_{n,k}^* - \hat{x}_{n,k+1}^* \right|}{N} \tag{21}$$

其中, $\hat{x}_{n,k}^*$ 和 $\hat{x}_{n,k+1}^*$ 分别表示第 n 个任务的第 k 次和第 k+1 次得到的“真值”.

通过算法 1 可以得到每个工人的质量和任务重要性. 除此之外, 还可以得到每个任务的初始“真值”以备最终真值发现使用. 现有研究表明真值发现容易受到初始值的影响, 因此, 经过 NWIE 后处理后的 NATURE 能得到较好的噪音“真值”精度.

3.3 效用感知的自适应剪枝

在得到工人的质量和任务的重要性后, 需要在满足 LDP 的约束下, 同时考虑 LDP 下真值发现的总体误差裁减掉部分值, 从而提高最终的数据质量. 为此, 首先给出本文需要的剪枝问题的形式化定义并形式化优化问题. 接着, 本文在分析 LDP 下总体误差的基础上设计具有多项式时间复杂度的求解算法.

假设有 M 个工人和 N 个任务形成的大小为 $M \times N$ 的稀疏矩阵. 已知每个工人的质量 w_s 和任务的重要性 y_n, 本文剪枝得到一个子集, 该子集要能够同时保证工人的质量最高、任务的重要性最大以及保证所有任务都有值被留下.

一种可能的做法是按照工人质量或者任务的重要性贪心地裁剪值, 不过这种做法可能降低最终数据集的质量, 甚至是导致真值发现失败.

(1) 从剪枝掉的数据量来说, 若剪枝少的话, 则无法起到剪枝的效果, 剪枝太多可能导致剩下数据的信息量不够. 显然, 无论哪种情况, 都会降低最终得到的“真值”精度.

(2) 从剪枝掉的工人来说, 其可能是高质量的工人, 这显然会浪费信息.

(3) 从剪枝掉的任务值来说可能导致某些任务没有值. 某些任务可能被工人做的次数较少, 有可能这些值都会被剪枝掉, 将会导致无法得到相关任务的“真值”.

为此, 本文提出效用感知的自适应剪枝方法 (utility-aware adaptive pruning, UAP). 该方法的主要思想是根据工人质量和任务的重要性效用感知且自适应地剪枝掉一些值. 下面首先给出本文形式化的剪枝问题 (公式 (22)), 并证明该剪枝问题是 NP-hard 的, 接着设计了具有多项式时间复杂度的求解算法并分析复杂度, 最后给出算法流程.

(1) 问题形式化及 NP-hard 证明

$$\begin{cases} \arg\max\limits_{\tilde{I}} \sum\limits_{\tilde{I}_{sn}=1} (w_s + y_n) \\ \text{s.t.} \quad \begin{cases} \sum_{s=1}^{M} \tilde{I}_{sn} \geqslant 1, \ \forall n = 1,2,\ldots,N \\ \sum_{n=1}^{N} \tilde{I}_{sn} \geqslant 0, \ \forall s = 1,2,\ldots,M \end{cases} \end{cases} \tag{22}$$

其中, $\tilde{I}$ 表示剪枝后工人所做任务的指示矩阵, 如果剪枝后第 s 个工人做了第 n 个任务, 那么 $\tilde{I}_{sn} = 1$, 否则 $\tilde{I}_{sn} = 0$. 目标函数表示需要剪枝后的矩阵中工人的质量和任务的重要性最大, 以期获得较好的表示. 第 1 个约束条件表示每个任务至少要有一个工人的提交值, 第 2 个约束条件表示某个工人的所有值可能完全被剪枝掉, 也可能完全被保留.

定理 4. 公式 (22) 所示的剪枝问题是 NP-hard 的.

证明: 本文将公式 (22) 所示的剪枝问题规约为已知的加权集合覆盖问题 (weighted set cover problem, WSCP).

由于 WSCP 是 NP-hard 的, 便可以证明公式 (22) 所示的剪枝问题也是 NP-hard 的.

首先给出 WSCP 的判定问题. 假定希望下载到 k 个需要的科研软件, 这些软件的集合为 U. S 为科研软件的下载镜像, 包含第 i 个软件的下载站为 S_i, 该站点的下载速度为 v_i, 且 S_i 中能下载的软件集合包含在 U 中. 如何确定下载软件的站点集合才能覆盖到 U 中所有软件并使总的下载速度最大.

其次将公式 (22) 所示剪枝问题规约为 WSCP. 具体地讲, 待覆盖的 N 个任务可以看作是软件集合 U, 每个工人做的任务集合可以看做 S_i, 每个任务的重要性 y_n 可以看作是站点的下载速度 v_i. 由于, 公式 (22) 还需要考虑工人的质量, 因此, 上述剪枝问题可以看作是更复杂的 WSCP 问题. 公式 (22) 所述的剪枝问题也是 NP-hard 的.

(2) 多项式时间复杂度的算法设计

由于剪枝问题的 NP-hard 特性, 本文需要设计出具有多项式时间复杂度的算法. 根据文献 [26], 所有任务隐私保护前后的平均绝对误差如公式 (23):

$$MAE \approx \frac{1}{N}\sum_{n=1}^{N}\left(\frac{d_2-d_1}{\varepsilon}+\sqrt{\frac{2}{\pi}}\sigma_{\text{avg}}\right) \tag{23}$$

其中, N 表示该任务的数量, σ_{avg} 表示某个任务的平均误差, d_1 和 d_2 表示某个任务域值的最小值和最大值.

再者, 根据文献 [6], $\sigma_{\text{avg}} \propto \frac{1}{w_{\text{avg}}^2}$, 那么对于工人, 本文有:

$$MAE_u \approx \frac{1}{N}\sum_{n=1}^{N}\left(\frac{d_2-d_1}{\varepsilon}+\sqrt{\frac{2}{\pi}}\frac{\sum_{s=1}^{M}\frac{1}{w_s^2}}{M}\right) \tag{24}$$

由于 d_1 和 d_2 等都为常量, 那么:

$$MAE_u \propto \frac{\sum_{s=1}^{M}\frac{1}{w_s^2}}{M} \tag{25}$$

那么, 为了对工人数据进行剪枝, 我们有如下定理 5.

定理 5. 给定质量从大到小排序好的工人集合, 从最小质量的工人开始剪枝. 假如 K 为剩余的工人数量, 如果满足公式 (26), 则停止剪枝.

$$\sum_{s=1}^{K}\frac{1}{w_s^2} \geqslant K\frac{1}{w_{K+1}^2} \tag{26}$$

证明: 已知 $1 \leqslant K \leqslant M-1$, $w_1 \geqslant w_K \geqslant w_M$, 设辅助变量 $z_s = \frac{1}{w_s^2}$. 根据公式 (24), 前 K 个工人被保留时候的总体误差如公式 (27) 所示:

$$MAE_K \propto \frac{\sum_{s=1}^{K} z_s}{K} \tag{27}$$

如果第 K+1 个工人也被保留, 那么总体误差为:

$$MAE_{K+1} \propto \frac{\sum_{s=1}^{K} z_s + z_{K+1}}{K+1} \tag{28}$$

如果第 K+1 个工人不被剪枝, 也就是说加上该工人后误差变小, 那么令 $MAE_K \geqslant MAE_{K+1}$, 得到:

$$MAE_K - MAE_{K+1} = \frac{\sum_{s=1}^{K} z_s - z_{K+1}K}{K(K+1)} \tag{29}$$

由于 $K(K+1)>0$, 因此, 如果 $\sum_{s=1}^{K} z_s \geqslant z_{K+1}K$, 那么 $MAE_K \geqslant MAE_{K+1}$.

同理, 为了对任务的数据进行剪枝, 有定理 6.

定理 6. 给定质量和重要性从大到小排序后的工人和任务集合, 从最小重要性的任务开始剪枝, 每次剪枝从做了该任务最小质量的工人开始, 如果满足公式 (30), 则停止剪枝.

证明: 对于第 n 个任务, 做了该任务的集合为 U_n, 那么对该任务来说总的误差为:

$$MAE_n \approx \frac{1}{|U_n|}\sum_{s\in U_n}\left(\frac{d_2-d_1}{\varepsilon/|M_s|}+\sqrt{\frac{2}{\pi}}\frac{1}{w_s^2}\right) \tag{30}$$

假设集合中工人质量最小的工人编号为 l, 排除掉第 l 个工人的总误差为:

$$MAE_{n^-} \approx \frac{1}{|U_n|-1}\sum_{s\in U_n\setminus\{u_l\}}\left(\frac{d_2-d_1}{\varepsilon/|M_s|}+\sqrt{\frac{2}{\pi}}\frac{1}{w_s^2}\right) \tag{31}$$

如果第 l 个值不被剪枝, 也就是说加上该值后误差变小, 那么令 $MAE_n \geqslant MAE_{n^-}$, 同时引入辅助变量 $z_s=\dfrac{d_2-d_1}{\varepsilon/|M_s|}+\sqrt{\dfrac{2}{\pi}}\dfrac{1}{w_s^2}$, 那么:

$$\begin{aligned}MAE_n-MAE_{n^-}&=\frac{1}{|U_n|}\sum_{s\in U_n}z_s-\frac{1}{|U_n|-1}\sum_{s\in U_n\setminus\{u_l\}}z_s\\&=\frac{|U_n|-1}{|U_n|(|U_n|-1)}\sum_{s\in U_n}z_s-\frac{|U_n|}{|U_n|(|U_n|-1)}\sum_{s\in U_n\setminus\{u_l\}}z_s\\&\propto(|U_n|-1)\sum_{s\in U_n}z_s-|U_n|\sum_{s\in U_n\setminus\{u_l\}}z_s=|U_n|z_l-\sum_{s\in U_n}z_s\\&=\frac{1}{|U_n|}\sum_{s\in U_n}\left(\frac{(d_2-d_1)(|M_l|-|M_s|)}{\varepsilon}+\sqrt{\frac{2}{\pi}}\left(\frac{1}{w_l^2}-\frac{1}{w_s^2}\right)\right)\end{aligned} \tag{32}$$

因此, 如果

$$\frac{1}{|U_n|}\sum_{s\in U_n}\left(\frac{(d_2-d_1)(|M_l|-|M_s|)}{\varepsilon}+\sqrt{\frac{2}{\pi}}\left(\frac{1}{w_l^2}-\frac{1}{w_s^2}\right)\right)\geqslant 0 \tag{33}$$

那么就停止剪枝.

(3) 自适应剪枝 (UAP) 算法流程

基于以上剪枝条件得到如下算法 2 流程. 首先进行初始化 (第 1, 2 行); 其次, 为每个任务分配至少一个值 (第 3–6 行); 再者, 优先给重要任务分配高质量的工人, 能提高真值发现的效用; 最后分别对工人 (第 7–9 行) 和任务 (第 10–12 行) 进行剪枝. 特别地, 由于某个任务不存在对应的提交值的话, 那么便无法得到该任务对应的“真值”, 进而导致真值发现失败. 因此采用本文算法能避免真值发现失败.

算法 2. 自适应剪枝 (UAP) 算法.

输入: 任务重要性矩阵 Y; 数据矩阵 X; 工人质量矩阵 W;
输出: X.

1) 对工人的质量 W 由大到小排序并对应更新 X;
2) 对任务的重要性 Y 由大到小排序并对应更新 X;
3) for (n=1 to N) do
4) 对任务 t_n 从 X 中从上到下选择最近一个值;
5) 在 X 中保留相应的值;
6) end for
7) for (s=1 to M) do

8) 根据定理 5 进行剪枝;
9) end for
10) for (n=1 to N) do
11) 根据定理 6 进行剪枝;
12) end for

通过对算法流程的分析, 可以发现算法主要包括初始化、为任务挑选值和对工人和任务进行剪枝. 对工人和任务分别进行初始化消耗 O(M) 和 O(N); 为每个任务挑选值并更新 X 消耗 O(N); 对每个工人和任务剪枝最多消耗 O(M^2) 和 O(N^2). 那么总的复杂度为 O(M^2+N^2+M+N).

4 算法分析与讨论

4.1 算法分析

在本节中, 首先进行算法分析, 包含隐私、效用和复杂度, 接着进行算法讨论.

4.1.1 隐私分析

定理 7. NATURE 算法严格满足 ε-LDP.

证明: 在 NATURE 中, 共包含 4 个阶段, 只在算法的阶段 1 接触原始数据, 假设每个工人 s 做的任务数用 $|M_s|$ 表示, 那么该工人对于每个任务 t 的隐私预算为 $\varepsilon_t=\varepsilon/|M_s|$, 根据定理 1, 该部分严格满足 ε-LDP.

在 NATURE 的后续阶段, 由于不接触原始数据, 因此不消耗隐私预算. 根据定理 3, 该部分也严格满足 ε-LDP. 根据定理 2, NATURE 对所有工人的数据严格满足 ε-LDP.

4.1.2 效用分析

首先在引理 1 中说明剪枝的必要性, 接着分析剪枝后的效果为什么更好.

引理 1. 剪枝的必要性.

证明: 只要证明剪枝后总误差更小, 便可证明剪枝的必要性.

根据文献 [6], LDP 下假设第 i 个工人的误差为 σ_i^2, 那么使用该工人数据引入的总误差为 $w_i^2\sigma_i^2$. 假设用 L_1 和 L_2 分别表示有某个工人和没有某个工人的总误差, 那么我们只需要证明 $L_1>L_2$ 即可.

$$L_1-L_2=\sum_{i=1}^{M}w_{ai}^2\sigma_i^2-\sum_{i=1}^{M-1}w_{bi}^2\sigma_i^2 \tag{34}$$

其中, w_{ai} 和 w_{bi} 分别表示第 i 个工人在有第 M 个工人数据和没有第 M 个工人数据时候的质量. 第 M 个工人的质量很小.

显然, 如果 $L_1^{-1}\geqslant L_2$, 其中 $L_1^{-1}=\sum_{i=1}^{M-1}w_{ai}^2\sigma_i^2$, 那么 $L_1\geqslant L_2$. 构造如下优化问题:

$$\begin{cases}\min\limits_{w_{ai},w_{bi}}\sum\limits_{i=1}^{M-1}\sigma_i^2(w_{ai}^2-w_{bi}^2)\\ \text{s.t.}\begin{cases}\sum\limits_{i=1}^{M}w_{ai}=1\\ \sum\limits_{i=1}^{M-1}w_{bi}=1\\ w_M\to 0\end{cases}\end{cases} \tag{35}$$

构造拉格朗日乘数可得:

$$L=\sum_{i=1}^{M-1}\sigma_i^2(w_{ai}^2-w_{bi}^2)+\lambda\left(\sum_{i=1}^{M}w_{ai}-1\right)+\mu\left(\sum_{i=1}^{M-1}w_{bi}-1\right) \tag{36}$$

对待求解变量求偏导可得:

$$\begin{cases} \dfrac{\partial L}{\partial w_{a1}} = 2\sigma_1^2 w_{a1} + \lambda \\ \dfrac{\partial L}{\partial w_{a2}} = 2\sigma_2^2 w_{a2} + \lambda \\ \quad \vdots \\ \dfrac{\partial L}{\partial w_{aM}} = \lambda \end{cases} \tag{37}$$

$$\begin{cases} \dfrac{\partial L}{\partial w_{b1}} = 2\sigma_1^2 w_{b1} + \mu \\ \dfrac{\partial L}{\partial w_{b2}} = 2\sigma_2^2 w_{b2} + \mu \\ \quad \vdots \\ \dfrac{\partial L}{\partial w_{bM-1}} = 2\sigma_{M-1}^2 w_{bM-1} + \mu \end{cases} \tag{38}$$

将公式 (37) 和公式 (38) 左右两边分别相加可得:

$$\sum_{i=1}^{M} \frac{\partial L}{\partial w_{ai}} + \sum_{i=1}^{M-1} \frac{\partial L}{\partial w_{bi}} = 2\sum_{i=1}^{M-1} \sigma_i^2 (w_{ai} - w_{bi}) + M\lambda + (M-1)\mu \tag{39}$$

根据拉格朗日乘子法, 需要 $\sum_{i=1}^{M} \frac{\partial L}{\partial w_{ai}} + \sum_{i=1}^{M-1} \frac{\partial L}{\partial w_{bi}} = 0$, 则要求让尽可能多的项为 0, 那么使得 $M\lambda + (M-1)\mu = 0$.根据前两个约束条件可得 $\sum_{i=1}^{M} w_{ai} - \sum_{i=1}^{M-1} w_{bi} = 0$, 进而有 $\sum_{i=1}^{M-1} (w_{ai} - w_{bi}) = -w_{aM}$. 显然如果

$$2\sum_{i=1}^{M-1} \sigma_i^2 (w_{ai} - w_{bi}) \geqslant 2\min\left(\sigma_i^2\right) \sum_{i=1}^{M-1} (w_{ai} - w_{bi}) \to 0 \tag{40}$$

便可证得结论.

由于 $w_M \to 0$, 需要使得 $-2\sum_{i=1}^{M-1} \sigma_i^2 w_{aM}$ 尽可能靠近 0.

显然当 $\sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \ldots = \sigma_{M-1}^2 = \min \sigma_i^2$ 时满足约束条件, 也就是 $L_1^{-1} \geqslant L_2$.

即证得 $L_1 > L_2$. 因此剪枝很有必要.

定理 8. NATURE 引入的总误差较小.

证明: 根据分析算法总的误差为:

$$MAE \approx \frac{1}{N} \sum_{n=1}^{N} \left(\frac{d_{2n} - d_{1n}}{\varepsilon} + \sqrt{\frac{2}{\pi}} \frac{\sum_{s=1}^{M} \frac{1}{w_s^2}}{M} \right) \tag{41}$$

通过公式 (41) 可以看出, 在变量中, 由于 $0 \leqslant w_s \leqslant 1$ 且 $M \in \{1,2,3,\ldots\}$, 因此 MAE 和某个工人的质量 w_s 和用户数量 M 负相关. 在 NATURE 中, 本文剪枝掉了低信息量的工人和任务值, 因此会得到较高的数据效用.

此外, 从引理 1 中达到最小误差的约束条件可以看出, 当最终数据的质量尽可能相等时误差最小. 本文通过对工人效用感知地进行剪枝, 可以保证留下的都是高质量的工人, 这些工人的质量接近, 因此总误差也会较小.

4.1.3 复杂度分析

本节将分析 NATURE 算法的通信复杂度和计算复杂度.

(1) 通信复杂度

通信仅发生在算法的第 1 阶段, 也就是每个工人上传自身加噪数据给服务器. 共需交换 O(MN). 显然该复杂

度和非隐私下的通信复杂度完全一致, 并没有增加多余的通信开销.

(2) 计算复杂度

从算法概述可以看出, 算法新增的主要时间消耗在于调用 NWIE 和 UAP 上. 对于 NWIE 来说, 其迭代求解工人权重、任务真值和任务重要性, 总的计算复杂度为 3O(MN).

对于 UAP, 根据第 3.3 节, 其计算开销为 O(M^2+N^2+M+N). 相对于非隐私的真值发现来说增加了 O($M^2+N^2+MN+M+N$). 不过, 实际中尤其发起者的招募预算限制, 工人或任务的数量往往较小, 同时由于数据的稀疏性, 工人所做任务的总记录数较少. 此外可以发现算法的主要计算在服务端进行, 本文可以合理地假设服务器有足够的计算能力, 而且也能利用分布式计算来进一步降低时间开销.

4.2 算法讨论

在 NATURE 中, 本文尽可能地提升了参与最终真值发现的数据的效用, 同时 NWIE 方法不仅可以估算工人的质量, 还可以为最终的真值发现提供初始化, 从而提升噪音“真值”的精度和真值发现算法的效率, 具体实验结果如第 5.4 节所示.

(1) LDP 下剪枝的合理性: 直观上来看, 剪枝后能使得留下的数据质量更高. 从引理 1 也可以看出当工人质量尽量相等时总误差最小, 因此在 LDP 下剪枝掉低质量的数据, 从而降低整体数据中蕴含的误差, 提升 LDP 下得到“真值”的精度.

(2) 剪枝顺序的确定: 本文先进行工人数据剪枝, 再进行任务剪枝, 其原因在于进行任务剪枝时候需要预先确定工人数量. 虽然, 先进行任务剪枝也需要确定工人数量, 但显然如果某个工人本身属于低质量工人, 其对“真值”精度的负面影响要大于某个任务的值. 因此必须先进行工人剪枝.

5 实验评估

5.1 数据集

本文采用两个公开可用的数据集: Population 和 Forecast[6,7].

(1) Population 数据集: 简称为 Pop, 该数据集收集自 2 415 个网站内关于 1 148 个城市的人口统计信息, 以美国人口局的数据作为“真值”. 由于本文关注在隐私和效用的权衡, 不失一般性, 本文随机选择 2 000 个网站关于 10 个城市的统计数据, 共计 16 816 条数据.

(2) Forecast 数据集: 简称为 For, 该数据集收集自 3 353 个气象观测点关于华中某地区的天气数据, 以中央气象台公布的数据为“真值”. 本文具体选择温度、湿度. 风速和风向这 4 个连续属性进行实验. 共计 12 427 条数据.

(3) 合成数据集: 简称为 Syn, 生成 1 000 个工人对 50 个任务的数据, 值的范围为 [1, 20], 设置真值为 10.5. 不失一般性, 对每个工人注入 $N(10.5,\sigma_i^2)$ 的噪音, 其中 σ_i 决定工人的质量, 本文定义 4 种类型的工人, 其方差分别为 1、5、10 和 50. 其比例分别为 20%、50%、20% 以及 10%.

5.2 评价标准与实验环境

为了充分评估所提算法和各个子方法的精度和收敛速度等, 本文采用 *MAE Change*、*KL-Divergence* 和目标函数值 J 来进行评估.

本文采用通用的 *MAE Change* 来判断最终得到的噪音“真值”的效用, 如下所示:

$$MAE\ Change=\frac{\sum_{t=1}^{N}\left|x_t^{*\prime}-x_t\right|-\sum_{t=1}^{N}\left|x_t^{*}-x_t\right|}{|T|} \tag{42}$$

其中, $|T|$ 是任务数, $x_t^{*\prime}$ 是第 t 个任务的隐私保护后的真值, x_t^* 是非隐私保护下得到的真值, x_t 是第 t 个任务真值的标准值. *MAE Change* 表示隐私保护前后 *MAE* (mean absolute error) 的变化量, 其值越小表示隐私保护对效用的影响越小.

为了评估隐私保护前后权重分布变化, 采用 *KL-Divergence* 来衡量隐私保护前后的权重变化, 其中 w 和 w' 分别是隐私保护前后的权重.

$$KL(w\|w') = \sum p(w)\log\frac{p(w)}{q(w')} \tag{43}$$

为了进一步评估初始值对结果的影响, 求得收敛时候的迭代次数, 使用如下收敛目标值 J[7]:

$$J = \sum_{s=1}^{M}\sum_{n=1}^{N} w_s(x_n^{s\prime} - \hat{x}_n^*)^2 \tag{44}$$

使用 Apple MacBook Pro 搭建所有的实验环境, 硬件信息为 MacOS monterey 系统、Apple M1 Pro、32 GB 内存、1 TB 固态硬盘. 所有实验采用 Python 3.7 来编写. 特别地, 迭代停止阈值 $\tau = 0.001$, 默认 $\varepsilon = 0.5$.

5.3 对比方法

根据对相关工作的分析, 本文发现现有方法无法直接用来解决本文所提出的研究问题, 为了验证所提方法的有效性, 将其与以下最新成果进行对比.

(1) VarFil: 该方法首先对工人的感知数据进行加噪, 然后根据噪音数据清除掉所有感知数据中方差位于 3σ 之外的值 (variance filtering, VarFil), 然后用剩下的数据进行真值发现.

(2) TLayer[23]: 该方法首先从一个预定义的分布中采样一个值作为方差, 然后将该方差作为均值为 0 的高斯分布中的方差并从该分布中对原始数据加噪, 最后服务器运行非隐私的真值发现算法来得到每个任务的真值, 该方法满足的是 (ε,δ)-LDP.

(3) PairsTD[25]: 该方法以 TLayer 方法为基础, 在此基础上聚合激励来降低上传值中的噪音量, 最后服务器运行非隐私的真值发现算法来得到每个任务的真值, 该方法满足的是 KL-LDP.

(4) PrivTDSI[26]: 该方法首先采样一部分数据, 然后对这部分数据进行加噪并上传, 根据上传后的加噪值推断未被采样到的值, 最后服务器运行非隐私的真值发现算法来得到每个任务的真值, 该方法严格满足 LDP.

(5) TESLA[27]: 该方法根据注入的噪音类型设计了噪音的过滤机制, 根据过滤后的噪音值进行非隐私的真值发现, 该方法严格满足 LDP.

5.4 对比实验

(1) ε 的影响

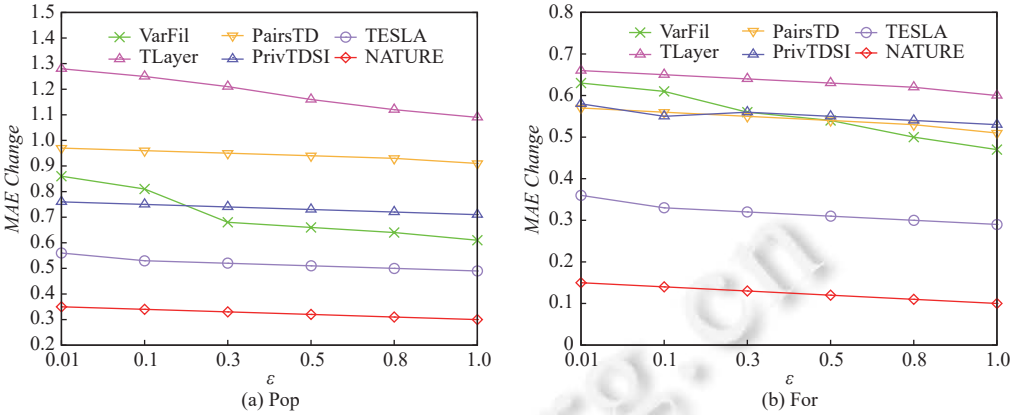
图 3 所示是 NATURE 和对比算法在不同隐私预算下的效果对比, 图 3(a) 和图 3(b) 分别对应 Pop 和 For 数据集. 可以看出, 随着 ε 的增加, NATURE 算法始终表现最好, 相对于最新的 TESLA 算法, 至少提升 20% 的精度. 这是因为在 NATURE 算法中, 本文预先自适应地剪枝掉了数据集中的异常值信息, 保留下来的是高质量信息. 显然在高质量信息上进行真值发现能得到精度较高的噪音“真值”. 尽管 VarFil 根据噪音方差去掉了一定的大的带噪值, 从而一定程度了抑制了误差传播. 然而一方面这些值可能来自稀疏任务, 进而导致真值发现失败, 另一方面他们可能对最终的“真值”有较大的正面影响, 直接剔除会降低“真值”的效用. 在其他对比算法中, 尽管研究者们采用了一系列方法来提升噪音“真值”的精度, 但是这些研究建立在数据是完全纯净的基础上或者预先在非隐私下对数据进行了预处理, 因此在实际中的表现较差. 特别地, TLayer 和 PairsTD 仅满足弱化的 LDP 保障, 即便是在这样的隐私放松下, NATURE 算法的精度仍然显著高于他们.

(2) 对权重分布的影响

由于权重在真值发现过程中处于主导地位, 在这部分的实验中, 本文验证所求得的噪音权重分布和真实权重分布的差距, 实验结果如图 4 所示. 图 4(a) 和图 4(b) 展示了 *KL-Divergence* 的结果, 图 4(c) 和图 4(d) 是在数据集中随机挑选的 8 个工人的真实权重值和噪音权重值的对比情况, 横坐标 ID 代表工人的编号, 纵坐标代表工人的权重值. 再者, TW (true weight) 代表非隐私下的真实权重, EW (estimated weight) 代表本文算法估计得到的权重.

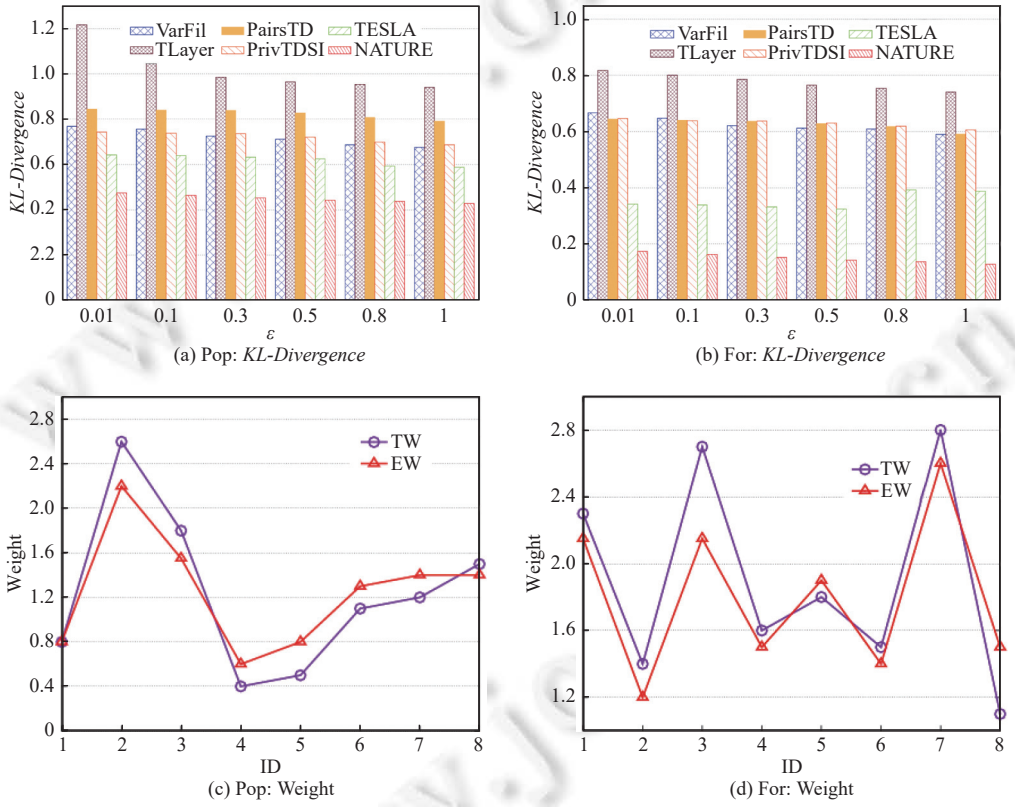
如图 4(a) 和图 4(b) 所示, 根据 NATURE 求得的权重分布和真实分布的 *KL-Divergence* 一直较小, 这说明根

据 NATURE 算法求得的权重分布距真实分布较为接近. 这是因为根据高质量的数据求得的噪音分布中蕴含的噪音规模较小. 特别地, 图 4(c) 和图 4(d) 中的实验结果也证明了这一现象.



(a) Pop (b) For

图 3 ε 的影响



(a) Pop: *KL-Divergence* (b) For: *KL-Divergence*

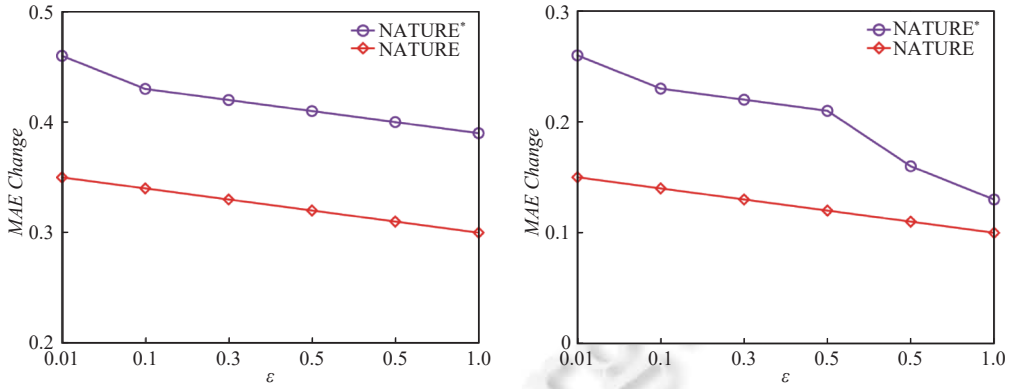
(c) Pop: Weight (d) For: Weight

图 4 对权重分布的影响

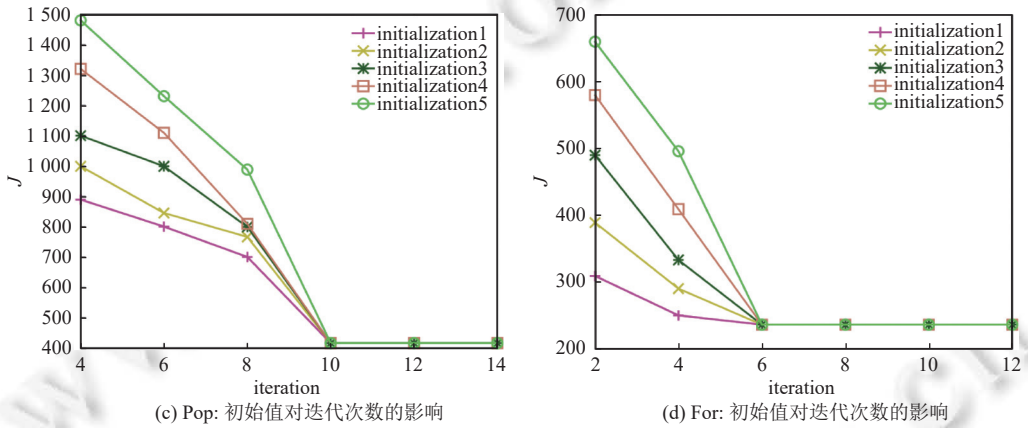
(3) 初始值的影响

为了验证在 LDP 约束下, 算法初始值对最终结果的影响, 本文做了如下实验. 第一, 本文进行在不采用 NATURE 中初始值对 CRH 进行初始化和进行初始化的效果对比, NATURE*表示不进行初始化的实验结果, 实验结果如图 5(a) 和图 5(b) 所示. 第二, 对 NATURE 中随机的初始权重分布分别扩大 1.1–1.5 倍, 算法分别对应 initalization1–initalization5, 实验结果如图 5(c) 和图 5(d) 以及表 2 所示.

从图 5(a) 和图 5(b) 中可以看出, 进行初始化的 NATURE 算法的效果显著好于不进行初始化的 NATURE*算法, 精度提升 10% 左右. 从图 5(c) 和图 5(d) 中可以看出, 在不同的算法初始化下, NATURE 算法仍保持较为稳定的迭代次数. 如表 2 所示, 在不采用初始化下, 算法的迭代次数有所提高, 这验证了初始化的作用, 也与非隐私下的发现一致.



(a) Pop: 初始值对精度的影响　　(b) For: 初始值对精度的影响



(c) Pop: 初始值对迭代次数的影响　　(d) For: 初始值对迭代次数的影响

图 5　初始值的影响

表 2　初始值对迭代次数的影响

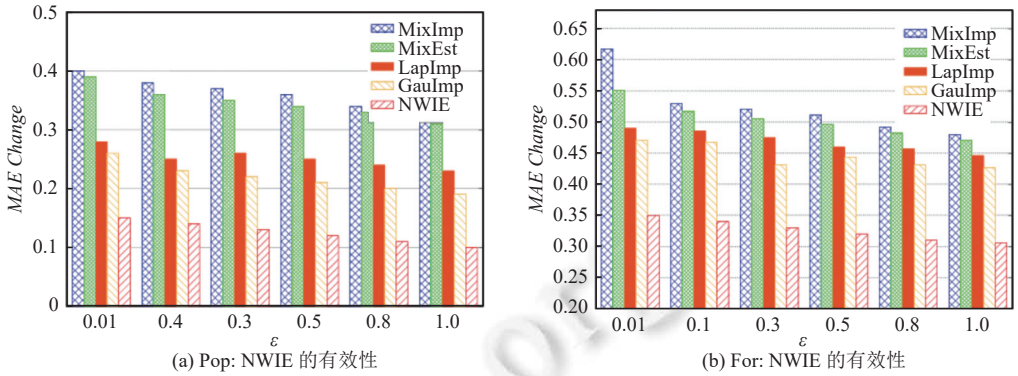
数据集	不初始化	迭代次数
Pop	initialization1*	12
	initialization2*	14
	initialization3*	13
	initialization4*	14
	initialization5*	12
For	initialization1*	9
	initialization2*	7
	initialization3*	8
	initialization4*	7
	initialization5*	8

注: *表示不采用第1阶段结果初始化下的迭代次数, 也就是随机初始化

(4) NWIE 有效性

NWIE 主要包含根据内源性的高斯噪音和注入的拉普拉斯噪音来同时估算工人的质量和任务的重要性, 因此

为了验证 NWIE 的效果, 本文将其与如下设计的对比算法进行对比. (1) MixImp (mixed importance): 该算法在考虑两种噪音的情况下, 仅估计任务的重要性. (2) MixEst (mixed estimation): 该算法在考虑两种噪音的情况下, 仅估计工人的权重. (3) LapImp (Laplace importance): 该算法在仅考虑注入的拉普拉斯的情况下, 同时估计工人的权重和任务的重要性. (4) GauImp (Gaussian importance): 该算法在仅考虑注入的拉普拉斯的情况下, 同时估计工人的权重和任务的重要性. 实验结果如图 6 所示.



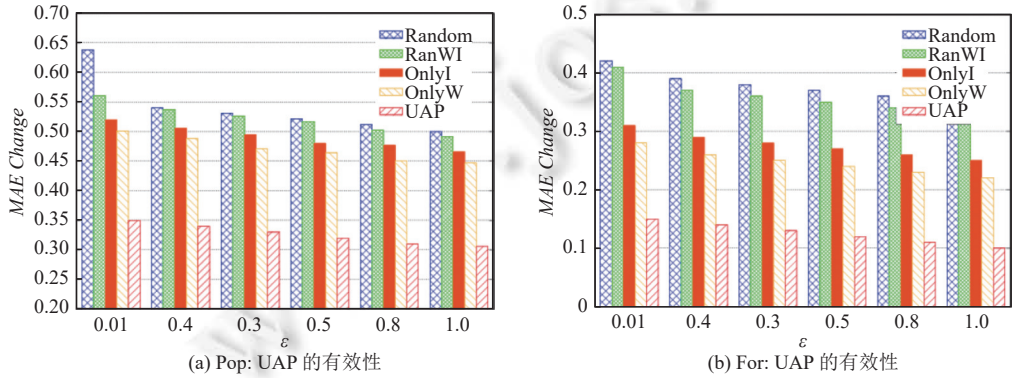
(a) Pop: NWIE 的有效性 (b) For: NWIE 的有效性

图 6 NWIE 有效性

从图 6 中可以看出, (1) MixEst 的效果好于 MixImp, 这说明在 LDP 的约束下, 工人的权重的影响大于工人重要性的影响, 这是因为在真值发现的过程中, 工人的权重处于主导地位. (2) LapImp 和 GauImp 的效果好于 MixImp 和 MixEst, 这说明同时考虑工人的权重和任务的重要性相对于仅考虑工人的权重和任务的重要性对精度的提升具有更大的贡献. 这是因为在 NATURE 中需要同时根据工人的权重和任务的重要性来对噪音数据进行剪枝. (3) GauImp 的效果好于 LapImp, 这说明采用欧氏距离来对噪音数据进行建模的效果好于曼哈顿距离. 这是因为, 在欧氏距离中采用平方项来对偏差进行建模, 会对偏差按平方放大, 显然其建模效果要好于绝对值项. (4) NWIE 的效果一直最好. 这是因为 NWIE 具有上述所有优点.

(5) UAP 有效性

UAP 主要包括根据工人质量和任务重要性分别进行效用感知自适应剪枝. 因此为了验证 UAP 的效果, 本文将其与如下设计的对比算法进行对比. (1) Random: 该算法随机剪枝 10% 的工人和任务数据. (2) RanWI (random weight and importance): 该算法根据工人权重和任务重要性随机剪枝 10% 的工人和任务数据. (3) OnlyI (only importance): 该算法仅根据任务的重要性对任务数据进行效用感知的自适应剪枝. (4) OnlyW (only weight): 该算法仅根据工人的权重对工人数据进行效用感知的自适应剪枝. 实验结果如图 7 所示.



(a) Pop: UAP 的有效性 (b) For: UAP 的有效性

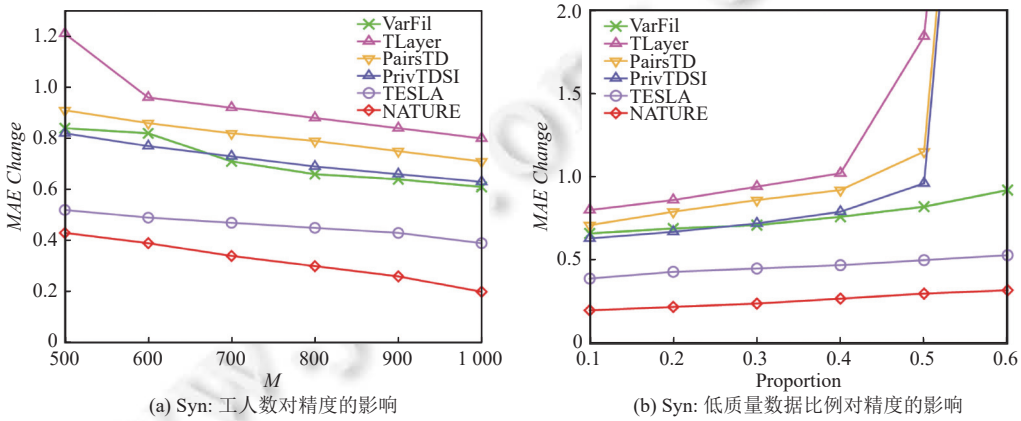
图 7 UAP 有效性

从图 7 中可以看出: (1) RanWI 的效果好于 Random, 这凸显了考虑工人质量和任务重要性的重要意义, 这是因为考虑工人质量和任务重要性能保留相对高质量的数据. (2) OnlyI 和 OnlyW 的效果均好于 Random 和 RanWI, 这是因为结合效用感知的自适应剪枝能使得保留下来的噪音数据能蕴含较小的噪音量, 这凸显了结合效用感知进行自适应剪枝的必要性. (3) OnlyW 的效果好于 OnlyI, 这说明在 LDP 的约束下, 工人权重的影响大于工人重要性的影响, 这是因为在真值发现的过程中, 工人的权重处于主导地位. (4) UAP 的效果一直最好. 这是因为 UAP 具有上述所有优点.

特别地, 从 UAP 和 NWIE 的对比来看, 不考虑 UAP 的 *MAE Change* (OnlyW) 要差于不考虑 NWIE 的 *MAE Change* (GauImp), 这说明进行剪枝的影响要大于进行权重和重要性估计, 从而与本文的研究动机相呼应.

(6) 合成数据集实验结果

为了验证 NATURE 在不同参数情况下的表现, 本文做了如下对比实验, 实验结果如图 8 所示.



(a) Syn: 工人数对精度的影响　　(b) Syn: 低质量数据比例对精度的影响

图 8　合成数据集实验结果

图 8(a) 展示了 NATURE 与对比算法在不同工人数量下的表现, 从图 8(a) 中可以看出, 随着工人数据量的增加, 所有算法的效果变好, 这是因为更多的工人会提供更多的有用信息. 此外, NATURE 算法的效果一直好于对比算法, 且在数据量越多的时候这种效果更加明显, 这是因为结合本文的 UAP 提前剪枝掉了垃圾信息, 保留下来了更多高质量的信息, NATURE 能充分利用数据中越来越多的高质量信息, 而对比算法在得益于越来越多的数据中的高质量信息同时, 也会被其中蕴含的异常值降低精度.

图 8(b) 展示了数据中不同低质量信息比例的结果, 本文将低质量信息的比例从 0.1 变更到 0.6. 从图 8(b) 中可以看出, 随着低质量信息比例的增加, 所有算法的效果变差, 这是因为数据中的有用信息变少. 此外, 当垃圾信息的比例超过一定值时比如 0.5, 除 TESLA 以外的对比算法的效果会显著变差, 其 *MAE Change* 显著增加 1.7 以上, 变得不可接受, 这是因为这些对比算法没有对噪音数据进行任何处理, 而 TESLA 和 NATURE 对数据进行了一定的处理. 再者, NATURE 的效果高于 TESLA, 这是因为 TESLA 对噪音数据进行了一定的提纯处理, 但是即便再怎么处理也是低质量的异常值数据, 仍然会对最终结果的精度有不良影响. 而 TESLA 是直接剪枝掉了这些低质量的异常值数据, 因此效果更好.

表 3 展示了不同数据分布下的实验结果. Ran 表示数据随机分布、Lap 和 Gau 分别表示数据按照多元拉普拉斯分布和多元高斯分布进行生成. 从表 3 中可以看出, 数据分布对最终“真值”精度的影响不大并且所提的 NATURE 算法的效果仍然最好, 这是因为在真值发现场景下, 发起者根据每个工人提交的数据得到的是每个任务的对应“真值”, 同时根据工人提交数据来估算工人的质量. 因为工人质量和对应的数据是一一对应的, 所以对最终的“真值”精度影响不大.

表 3 Syn: 不同数据分布的鲁棒性

分布	对比算法	*MAE Change*
Ran	VarFil	0.6471
	TLayer	0.8322
	PairsTD	0.7636
	PrivTDSI	0.6321
	TESLA	0.4555
	NATURE	0.2434
Lap	VarFil	0.6446
	TLayer	0.8308
	PairsTD	0.7612
	PrivTDSI	0.6289
	TESLA	0.4531
	NATURE	0.2420
Gau	VarFil	0.6458
	TLayer	0.8313
	PairsTD	0.7625
	PrivTDSI	0.6303
	TESLA	0.4542
	NATURE	0.2424

6 结束语

本文针对连续值场景, 基于本地差分隐私研究了目前已有研究工作中尚未充分考虑的含异常值的真值发现问题, 并针对此问题提出 NATURE 算法, 然后从理论上分析了算法的隐私、效用和复杂度, 最后在两个真实数据集和一个合成数据集上验证了算法的有效性. 在有些场景下, 工人提交数据可能是离散值, 甚至是文本、图像等, 在未来研究中, 拟针对上述场景下高效率且高效用的本地差分隐私的真值发现问题做进一步探索.

References:

[1] Xiong JB, Bi RW, Tian YL, Liu XM, Ma JF. Security and privacy in mobile crowdsensing: Models, progresses, and trends. Chinese Journal of Computers, 2021, 44(9): 1949–1966 (in Chinese with English abstract). [doi: 10.11897/SP.J.1016.2021.01949]

[2] Tong YX, Yuan Y, Cheng YR, Chen L, Wang GR. Survey on spatiotemporal crowdsourced data management techniques. Ruan Jian Xue Bao/Journal of Software, 2017, 28(1): 35–58 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/5140.htm [doi: 10.13328/j.cnki.jos.005140]

[3] Chai CL, Li GL, Zhao TY, Luo YY, Yu MH. Crowd-powered database system: A survey. Chinese Journal of Computers, 2020, 43(5): 948–972 (in Chinese with English abstract). [doi: 10.11897/SP.J.1016.2020.00948]

[4] Feng JH, Li GL, Feng JH. A survey on crowdsourcing. Chinese Journal of Computers, 2015, 38(9): 1713–1726 (in Chinese with English abstract). [doi: 10.11897/SP.J.1016.2015.01713]

[5] Ji SL, Du TY, Li JF, Shen C, Li B. Security and privacy of machine learning models: A survey. Ruan Jian Xue Bao/Journal of Software, 2021, 32(1): 41–67 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/6131.htm [doi: 10.13328/j.cnki.jos.006131]

[6] Li Q, Li YL, Gao J, Su L, Zhao B, Demirbas M, Fan W, Han JW. A confidence-aware approach for truth discovery on long-tail data. Proc. of the VLDB Endowment, 2014, 8(4): 425–436. [doi: 10.14778/2735496.2735505]

[7] Li Q, Li YL, Gao J, Zhao B, Fan W, Han JW. Resolving conflicts in heterogeneous data by truth discovery and source reliability estimation. In: Proc. of the 2014 ACM SIGMOD Int'l Conf. on Management of Data. Snowbird: ACM, 2014. 1187–1198. [doi: 10.1145/2588555.2610509]

[8] Zeng JR, Chen H, Peng H, Wu Y, Li CP, Wang S. Privacy preservation in mobile participatory sensing. Chinese Journal of Computers, 2016, 39(3): 595–614 (in Chinese with English abstract). [doi: 10.11897/SP.J.1016.2016.00595]

[9] Zheng YF, Duan HY, Yuan XL, Wang C. Privacy-aware and efficient mobile crowdsensing with truth discovery. IEEE Trans. on Dependable and Secure Computing, 2020, 17(1): 121–133. [doi: 10.1109/TDSC.2017.2753245]

[10] Ye QQ, Meng XF, Zhu MJ, Huo Z. Survey on local differential privacy. Ruan Jian Xue Bao/Journal of Software, 2018, 29(7): 1981–2005

(in Chinese with English abstract). http://www.jos.org.cn/1000-9825/5364.htm [doi: 10.13328/j.cnki.jos.005364]

[11] Erlingsson Ú, Pihur V, Korolova A. RAPPOR: Randomized aggregatable privacy-preserving ordinal response. In: Proc. of the 2014 ACM SIGSAC Conf. on Computer and Communications Security. Scottsdale: ACM, 2014. 1054–1067. [doi: 10.1145/2660267.2660348]

[12] Tang J, Korolova A, Bai XL, Wang XQ, Wang XF. Privacy loss in Apple's implementation of differential privacy on MacOS 10.12. arXiv:1709.02753, 2017.

[13] Ding BL, Kulkarni J, Yekhanin S. Collecting telemetry data privately. In: Proc. of the 31st Int'l Conf. on Neural Information Processing Systems. Long Beach: Curran Associates Inc., 2017. 3574–3583.

[14] Husain H, Balle B, Cranko Z, Nock R. Local differential privacy for sampling. In: Proc. of the 23rd Int'l Conf. on Artificial Intelligence and Statistics. Palermo: PMLR, 2020. 3404–3413.

[15] Lin DJ, Wu YD, Gan WS. Sybil-resistant truth discovery in crowdsourcing by exploiting the long-tail effect. In: Proc. of the 21st IEEE Int'l Conf. on Trust, Security and Privacy in Computing and Communications. Wuhan: IEEE, 2022. 1545–1550. [doi: 10.1109/TrustCom56396.2022.00221]

[16] Xu F, Sheng VS, Wang MW. A unified perspective for disinformation detection and truth discovery in social sensing: A survey. ACM Computing Surveys (CSUR), 2023, 55(1): 6. [doi: 10.1145/3477138]

[17] Wang D, Kaplan L, Abdelzaher TF. Maximum likelihood analysis of conflicting observations in social sensing. ACM Trans. on Sensor Networks (TOSN), 2014, 10(2): 30. [doi: 10.1145/2530289]

[18] Li HW, Zhao B, Fuxman A. The wisdom of minority: Discovering and targeting the right group of workers for crowdsourcing. In: Proc. of the 23rd Int'l Conf. on World Wide Web. Seoul: ACM, 2014. 165–176. [doi: 10.1145/2566486.2568033]

[19] Zhang D, Wang D, Vance N, Zhang Y, Mike S. On scalable and robust truth discovery in big data social media sensing applications. IEEE Trans. on Big Data, 2019, 5(2): 195–208. [doi: 10.1109/TBDATA.2018.2824812]

[20] Xu C, Jia Y, Zhu LH, Zhang C, Jin GX, Sharif K. TDFL: Truth discovery based Byzantine robust federated learning. IEEE Trans. on Parallel and Distributed Systems, 2022, 33(12): 4835–4848. [doi: 10.1109/TPDS.2022.3205714]

[21] Zhang HN, Li MH, Sun YB, Qu GQ. Robust truth discovery against multi-round data poisoning attacks. In: Proc. of the 17th Int'l Conf. on Wireless Algorithms, Systems, and Applications. Dalian: Springer, 2022. 258–270. [doi: 10.1007/978-3-031-19208-1_22]

[22] Sun HP, Dong BX, Wang H, Yu T, Qin Z. Truth inference on sparse crowdsourcing data with local differential privacy. In: Proc. of the 2018 IEEE Int'l Conf. on Big Data. Seattle: IEEE, 2018. 488–497. [doi: 10.1109/BigData.2018.8622635]

[23] Li YL, Xiao HP, Qin Z, Miao CL, Su L, Gao J, Ren K, Ding BL. Towards differentially private truth discovery for crowd sensing systems. In: Proc. of the 40th Int'l Conf. on Distributed Computing Systems. Singapore: IEEE, 2020. 1156–1166. [doi: 10.1109/ICDCS47774.2020.00037]

[24] Sun P, Wang ZB, Feng YH, Wu LT, Li YJ, Qi HR, Wang Z. Towards personalized privacy-preserving incentive for truth discovery in crowdsourced binary-choice question answering. In: Proc. of the 39th IEEE Conf. on Computer Communications. Toronto: IEEE, 2020. 1133–1142. [doi: 10.1109/INFOCOM41043.2020.9155429]

[25] Sun P, Wang ZB, Wu LT, Feng YH, Pang XY, Qi HR, Wang Z. Towards personalized privacy-preserving incentive for truth discovery in mobile crowdsensing systems. IEEE Trans. on Mobile Computing, 2022, 21(1): 352–365. [doi: 10.1109/TMC.2020.3003673]

[26] Zhang PF, Cheng X, Su S, Zhu BY. PrivTDSI: A local differentially private approach for truth discovery via sampling and inference. IEEE Trans. on Big Data, 2023, 9(2): 471–484. [doi: 10.1109/TBDATA.2022.3186175]

[27] Zhang PF, Cheng X, Su S, Wang N. Effective truth discovery under local differential privacy by leveraging noise-aware probabilistic estimation and fusion. Knowledge-based Systems, 2023, 261: 110213. [doi: 10.1016/j.knosys.2022.110213]

[28] Cao XY, Jia JY, Gong NZ. Data poisoning attacks to local differential privacy protocols. In: Proc. of the 30th USENIX Security Symp. USENIX, 2019. 947–964.

[29] Cheu A, Smith A, Ullman J. Manipulation attacks in local differential privacy. In: Proc. of the 2021 IEEE Symp. on Security and Privacy. San Francisco: IEEE, 2021. 883–900. [doi: 10.1109/SP40001.2021.00001]

[30] Li ZT, Zheng ZR, Guo SM, Guo B, Xiao F, Ren K. Disguised as privacy: Data poisoning attacks against differentially private crowdsensing systems. IEEE Trans. on Mobile Computing, 2023, 22(9): 5155–5169. [doi: 10.1109/TMC.2022.3173642]

[31] Zhang HN, Li MH. Multi-round data poisoning attack and defense against truth discovery in crowdsensing systems. In: Proc. of the 23rd IEEE Int'l Conf. on Mobile Data Management. Paphos: IEEE, 2022. 109–118. [doi: 10.1109/MDM55031.2022.00036]

[32] Song SR, Xu L, Zhu LH. Efficient defenses against output poisoning attacks on local differential privacy. IEEE Trans. on Information Forensics and Security, 2023, 18: 5506–5521. [doi: 10.1109/TIFS.2023.3305873]

[33] Huang K, Ouyang GY, Ye QQ, Hu HB, Zheng BL, Zhao X, Zhang RY, Zhou XF. LDPGuard: Defenses against data poisoning attacks to local differential privacy protocols. IEEE Trans. on Knowledge and Data Engineering, 2024, 36(7): 3195–3209. [doi: 10.1109/TKDE.

2024.3358909]
[34] Zheng ZR, Li ZT, Huang C, Long SQ, Li MS, Shen XM. Data poisoning attacks and defenses to LDP-based privacy-preserving crowdsensing. IEEE Trans. on Dependable and Secure Computing, 2024, 21(5): 4861–4878. [doi: 10.1109/TDSC.2024.3363507]
[35] Hossain MT, Islam S, Badsha S, Shen HT. DeSMP: Differential privacy-exploited stealthy model poisoning attacks in federated learning. In: Proc. of the 17th Int'l Conf. on Mobility, Sensing and Networking. Exeter: IEEE, 2021. 167–174. [doi: 10.1109/MSN53354.2021.00038]
[36] Zhong YS, Wang LP. PROFL: A privacy-preserving federated learning method with stringent defense against poisoning attacks. In: Proc. of the 27th Int'l Conf. on Computer Supported Cooperative Work in Design. Tianjin: IEEE, 2024. 260–265. [doi: 10.1109/CSCWD61410.2024.10580526]
[37] Huang YX, Yang G, Zhou H, Dai H, Yuan D, Yu S. VPPFL: A verifiable privacy-preserving federated learning scheme against poisoning attacks. Computers & Security, 2024, 136: 103562. [doi: 10.1016/j.cose.2023.103562]
[38] Zhang C, Zhu LH, Xu C, Liu XM, Sharif K. Reliable and privacy-preserving truth discovery for mobile crowdsensing systems. IEEE Trans. on Dependable and Secure Computing, 2021, 18(3): 1245–1260. [doi: 10.1109/TDSC.2019.2919517]
[39] Xue KP, Zhu B, Yang QY, Gai N, Wei DSL, Yu NH. InPPTD: A lightweight incentive-based privacy-preserving truth discovery for crowdsensing systems. IEEE Internet of Things Journal, 2021, 8(6): 4305–4316. [doi: 10.1109/JIOT.2020.3029294]
[40] Zhao JC, Zhu B, Li J, Yuan SX, Xue KP, Zhang XC. Privacy-preserving truth discovery with outlier detection in mobile crowdsensing systems. In: Proc. of the 2022 IEEE Global Communications Conf. Rio de Janeiro: IEEE, 2022. 4352–4357. [doi: 10.1109/GLOBECOM48099.2022.10000662]
[41] Wu HQ, Wang LM, Cheng K, Yang DJ, Tang J, Xue GL. Privacy-enhanced and practical truth discovery in two-server mobile crowdsensing. IEEE Trans. on Network Science and Engineering, 2022, 9(3): 1740–1755. [doi: 10.1109/TNSE.2022.3151228]
[42] Tang JC, Fu SJ, Liu XM, Luo YC, Xu M. Achieving privacy-preserving and lightweight truth discovery in mobile crowdsensing. IEEE Trans. on Knowledge and Data Engineering, 2022, 34(11): 5140–5153. [doi: 10.1109/TKDE.2021.3054409]
[43] Zhang C, Zhao MY, Zhu LH, Wu T, Liu XM. Enabling efficient and strong privacy-preserving truth discovery in mobile crowdsensing. IEEE Trans. on Information Forensics and Security, 2022, 17: 3569–3581. [doi: 10.1109/TIFS.2022.3207905]
[44] Liu XT, Zhou SW, Zhang W, Dong T, Li KQ. Privacy-preserving truth discovery for collaborative-cloud encryption in mobile crowdsensing. IEEE Systems Journal, 2023, 17(3): 4990–5001. [doi: 10.1109/JSYST.2023.3274812]
[45] Xiong P, Li GR, Liu HZ, Hu YY, Jin DW. Decentralized privacy-preserving truth discovery for crowd sensing. Information Sciences, 2023, 632: 730–741. [doi: 10.1016/j.ins.2023.03.046]
[46] Liu R, Pan JP. Lightweight privacy-preserving truth discovery for vehicular air quality monitoring. Digital Communications and Networks, 2023, 9(1): 280–291. [doi: 10.1016/j.dcan.2022.03.021]
[47] Li YL, Miao CL, Su L, Gao J, Li Q, Ding BL, Qin Z, Ren K. An efficient two-layer mechanism for privacy-preserving truth discovery. In: Proc. of the 24th ACM SIGKDD Int'l Conf. on Knowledge Discovery & Data Mining. London: ACM, 2018. 1705–1714. [doi: 10.1145/3219819.3219998]
[48] Wang D, Xu JH. Inferring ground truth from crowdsourced data under local attribute differential privacy. Theoretical Computer Science, 2021, 865: 85–98. [doi: 10.1016/j.tcs.2021.02.039]

附中文参考文献:

[1] 熊金波, 毕仁万, 田有亮, 刘西蒙, 马建峰. 移动群智感知安全与隐私: 模型、进展与趋势. 计算机学报, 2021, 44(9): 1949–1966. [doi: 10.11897/SP.J.1016.2021.01949]
[2] 童咏昕, 袁野, 成雨蓉, 陈雷, 王国仁. 时空众包数据管理技术研究综述. 软件学报, 2017, 28(1): 35–58. http://www.jos.org.cn/1000-9825/5140.htm [doi: 10.13328/j.cnki.jos.005140]
[3] 柴成亮, 李国良, 赵天宇, 骆昱宇, 于明鹤. 众包数据库综述. 计算机学报, 2020, 43(5): 948–972. [doi: 10.11897/SP.J.1016.2020.00948]
[4] 冯剑红, 李国良, 冯建华. 众包技术研究综述. 计算机学报, 2015, 38(9): 1713–1726. [doi: 10.11897/SP.J.1016.2015.01713]
[5] 纪守领, 杜天宇, 李进锋, 沈超, 李博. 机器学习模型安全与隐私研究综述. 软件学报, 2021, 32(1): 41–67. http://www.jos.org.cn/1000-9825/6131.htm [doi: 10.13328/j.cnki.jos.006131]
[8] 曾菊儒, 陈红, 彭辉, 吴垚, 李翠平, 王珊. 参与式感知隐私保护技术. 计算机学报, 2016, 39(3): 595–614. [doi: 10.11897/SP.J.1016.2016.00595]
[10] 叶青青, 孟小峰, 朱敏杰, 霍峥. 本地化差分隐私研究综述. 软件学报, 2018, 29(7): 1981–2005. http://www.jos.org.cn/1000-9825/5364.htm [doi: 10.13328/j.cnki.jos.005364]

张朋飞(1992－), 男, 博士, 讲师, CCF 专业会员, 主要研究领域为数据隐私保护, 可信人工智能.

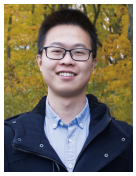
刘西蒙(1988－), 男, 博士, 教授, 博士生导师, CCF 高级会员, 主要研究领域为数据隐私保护, 数据安全.

朱伊波(1999－), 男, 硕士生, 主要研究领域为数据隐私保护.

孙笠(1994－), 男, 博士, 讲师, 主要研究领域为数据挖掘, 机器学习.

程祥(1984－), 男, 博士, 教授, 博士生导师, 主要研究领域为数据隐私保护, 可信人工智能.

方贤进(1970－), 男, 博士, 教授, 博士生导师, CCF 杰出会员, 主要研究领域为网络信息安全.



张治坤(1993－), 男, 博士, 副教授, 博士生导师, 主要研究领域为隐私计算, 数据隐私保护, 机器学习隐私与安全.

张吉(1977－), 男, 博士, 教授, 博士生导师, 主要研究领域为数据科学, 数据挖掘, 机器学习, 隐私保护.